



CK.0000069078

GỌC KIẾNG

THỐNG KÊ HỌC TRONG NGHIÊN CỨU KHOA HỌC

NGUYỄN
ĐC LỊCH



NHÀ XUẤT BẢN XÂY DỰNG



TS. NGUYỄN NGỌC KIẾNG

THỐNG KÊ HỌC TRONG NGHIÊN CỨU KHOA HỌC

(Tái bản)

NHÀ XUẤT BẢN XÂY DỰNG
HÀ NỘI - 2013

LỜI NÓI ĐẦU

Thống kê học không những giúp các nhà nghiên cứu biết cách tiến hành thu thập số liệu, xử lý số liệu, phân tích và lý giải kết quả nghiên cứu mà hơn thế nữa là giúp định hướng nghiên cứu và dự tính - dự báo các vấn đề khóa học một cách đúng đắn và hợp lý nhất. Nội dung quyển sách này bao gồm những vấn đề căn bản nhất của thống kê học như các phân phối trong thống kê học, thuật so sánh các nghiệm thức, tương quan và hồi quy, đặc biệt là phương pháp thiết lập các mô hình toán học... Thiết nghĩ những vấn đề được trình bày trong sách sẽ giúp ích nhiều cho các bạn sinh viên, nghiên cứu sinh cũng như các nhà khoa học đang làm công tác nghiên cứu khoa học. Trong quá trình biên soạn chắc không tránh khỏi những thiếu sót, tác giả kính mong bạn đọc nhiệt tình đóng góp ý kiến.

Tác giả

NGUYỄN NGỌC KIẾNG

Chương 1

CÁC ĐẠI LƯỢNG THỐNG KÊ

1.1. TRUNG BÌNH

Trong đa số các trường hợp, đặc biệt là các trường hợp mà các số liệu là kết quả của các “thăng giáng ngẫu nhiên” thì đại lượng trung bình cộng, thường gọi là trung bình, là đại lượng tiêu biểu và đặc trưng nhất được dùng làm đại biểu hay đại diện cho quần thể (population) hay mẫu (sample) lấy ra từ quần thể đó.

Nếu x_i là số liệu của các thăng giáng ngẫu nhiên được lấy ra từ một quần thể nào đó và n_i là số số liệu giống nhau (tần số) của mỗi trị số x_i , còn n là tổng số các số liệu được lấy ra (kích thước của mẫu) thì trung bình (cộng) của mẫu, ký hiệu là $\bar{x}$, được định nghĩa như sau:

$$\frac{\underbrace{x_1 + x_1 + \dots + x_1}_{n_1} + \underbrace{x_2 + x_2 + \dots + x_2}_{n_2} + \underbrace{x_K + x_K + \dots + x_K}_{n_K}}{n} \\ = \underbrace{\bar{x} + \bar{x} + \dots + \bar{x}}_n$$

Hay: $n_1 x_1 + n_2 x_2 + \dots + n_K x_K = n \bar{x}$ với $n = \sum_{i=1}^K n_i$

Hay:
$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^K n_i x_i$$

Bây giờ nếu như kích thước của mẫu bằng kích thước của quần thể, hay $n \approx \infty$, thì trung bình của mẫu ($\bar{x}$) trở thành trung bình của quần thể và thường được ký hiệu là m .

Trong nhiều trường hợp để đơn giản hóa việc tính toán nếu số liệu nhiều thì nên tiến hành phân nhóm số liệu, nếu giá trị của số liệu lớn hay phức tạp thì nên tiến hành mã hóa số liệu. Như vậy, nếu như chúng ta vừa tiến hành phân nhóm số liệu vừa tiến hành mã hóa số liệu thì công việc tính toán sẽ đơn giản đi rất nhiều.

Trong quá trình tiến hành phân nhóm số liệu cần chú ý rằng, số lượng nhóm là tùy chúng ta quyết định lấy, thường là 5 - 10 nhóm, từ đó chúng ta xác định được độ lớn của nhóm và trị trung tâm của mỗi nhóm là giá trị đại diện cho nhóm đó. Tần số của mỗi nhóm được xác định dựa trên khoảng nửa hờ [) hoặc (], riêng nhóm cuối cùng có thể là khoảng kín []. Chẳng hạn, nếu như biên độ dao động của các số liệu là $\omega = x_{\max} - x_{\min} = 23,8 - 10,1 = 13,7$, và nếu chúng ta dự định phân làm 7 nhóm thì độ lớn của mỗi nhóm sẽ là $b = 13,7/7 \approx 2$, và như vậy phân nhóm đầu tiên sẽ là [10; 12) với trị trung tâm làm đại diện cho nhóm này sẽ là $x_1 = 11$. Phân nhóm cuối cùng sẽ là [22; 24) với trị trung tâm làm đại diện cho nhóm này sẽ là $x_7 = 23$. Chú ý rằng, chúng ta có thể lấy trị $x_{\min}$ làm trị khởi đầu của phân nhóm đầu tiên nếu thấy tiện lợi hơn, chẳng hạn, nếu biên độ dao động của các số liệu là $\omega = 77 - 14 = 63$, và nếu như chúng ta dự định phân làm 7 nhóm thì độ lớn của mỗi nhóm sẽ là $b = 63/7 = 9$, và như vậy phân nhóm đầu tiên sẽ là [14; 23) với trị trung tâm làm đại diện cho nhóm này là $x_1 = 18,5$. Phân nhóm cuối cùng sẽ là khoảng kín [68; 77] với trị trung tâm làm đại diện cho nhóm này là $x_7 = 72,5$. Như đã nói ở trên, phân nhóm cuối cùng có thể là khoảng nửa hờ hoặc khoảng kín, chẳng hạn với $\omega = 21,9 - 8,4 = 13,5$ và nếu dự định chia làm 7 nhóm thì độ lớn của nhóm sẽ là $b = 13,5/7 \approx 2$, và như vậy phân nhóm đầu tiên sẽ là [8,4; 10,4) với $x_1 = 9,4$ còn phân nhóm cuối cùng sẽ là khoảng nửa hờ [20,4; 22,4) với $x_7 = 21,4$.

Trong khi mã hóa số liệu chúng ta cần nắm vững các kỹ thuật tính toán để tránh những sai lầm đáng tiếc có thể xảy ra. Trước tiên, chúng ta chọn một giá trị bất kỳ x_0 làm gốc, trị số gốc x_0 được chọn sao cho có lợi nhất về mặt tính toán, do đó, x_0 thường được chọn nằm lân cận trung điểm của đoạn $[x_{\min}; x_{\max}]$. Trong trường hợp số liệu không phân nhóm thì tiến trình mã hóa số liệu sẽ như sau : $u_i = x_i - x_0 \Leftrightarrow x_i = x_0 + u_i$, do đó $M[x_0 + u_i] = M[x_0] + M[u_i]$ hay $\bar{x} = x_0 + \bar{u}$. Còn trong trường hợp số liệu có phân nhóm thì tiến trình mã hóa số liệu sẽ như sau: $u_i = (x_i - x_0)/b \Leftrightarrow x_i = x_0 + bu_i$, do đó $M[x_i] = M[x_0 + bu_i] = M[x_0] + M[bu_i] = M[x_0] + bM[u_i]$ hay $\bar{x} = x_0 + b.\bar{u}$. Chú ý rằng, trong trường hợp số liệu không phân nhóm chúng ta cũng có thể mã hóa số liệu theo công thức $u_i = (x_i - x_0)/a$ với a là trị số bất kỳ khác 1 tùy chúng ta quyết định. Như vậy, trong mọi trường hợp chúng ta đều có thể tiến hành mã hóa số liệu theo công thức tổng quát sau đây: $u_i = \alpha(x_i - x_0)$ với α là trị số bất kỳ tùy ý chúng ta quyết định lựa chọn sao cho có lợi nhất về mặt tính toán, chẳng hạn, với trường hợp số liệu có phân nhóm chúng ta chọn

$\alpha = 1/b$ với b là độ lớn của nhóm. Vì rằng $u_i = \alpha(x_i - x_0) \Leftrightarrow x_i = x_0 + \frac{1}{\alpha}u_i$,

do đó $M[x_i] = M\left[x_0 + \frac{1}{\alpha}u_i\right] = M[x_0] + M\left[\frac{1}{\alpha}u_i\right] = M[x_0] + \frac{1}{\alpha}M[u_i]$ hay

$\bar{x} = x_0 + \frac{1}{\alpha}\bar{u}$. Công việc tính toán trung bình $\bar{u}$ của số liệu đã được mã hóa

chắc chắn sẽ đơn giản hơn rất nhiều so với việc tính toán trực tiếp trung bình $\bar{x}$ từ số liệu ban đầu. Các bảng tính toán số liệu sau đây (bảng 1.1 và 1.2) sẽ minh họa cho những điều trình bày ở trên.

Bảng 1.1. Tính trung bình $\bar{x}$ trong trường hợp không mã hóa và có mã hóa số liệu (với $x_0 = 75$ và $a = 1/\alpha = 5$)

x_i	n_i	$x_i n_i$	$u_i = (x_i - 75)/5$	$n_i u_i$
95	4	380	4	16
90	6	540	3	18
85	12	1020	2	24
80	19	1520	1	19
75	37	2775	0	0
70	24	1680	-1	-24
65	11	715	-2	-22
60	6	360	-3	-18
55	4	220	-4	-16
50	2	100	-5	-10
Tổng	125	9310		-13
$\bar{x} = 9310/125 = 74,48$			$\bar{x} = 75 + 5\left(\frac{-13}{125}\right) = 74,48$	

Bảng 1.2. Tính trung bình $\bar{x}$ trong trường hợp số liệu không phân nhóm và có phân nhóm (với $x_0 = 11,5$ và $b = 1$)

8,35	13,65	11,00	11,35	13,75	12,76	13,25
8,74	15,25	11,00	11,50	16,00	12,85	13,26
9,25	14,51	11,25	11,50	12,15	14,75	13,40
9,50	10,50	14,50	11,62	12,25	14,26	13,50
9,75	10,75	14,23	11,75	12,35	12,85	14,00
10,24	10,76	16,25	12,00	12,50	13,15	16,00
$\sum x_i = 522,23, n = 42, \bar{x} = \frac{522,23}{42} \approx 12,434$						

	x_i	n_i	$u_i = (x_i - 11,5)/1$	$n_i u_i$
8 - 9	8,5	2	-3	-6
9 - 10	9,5	3	-2	-6
10 - 11	10,5	4	-1	-4
11 - 12	11,5	8	0	0
12 - 13	12,5	8	1	8
13 - 14	13,5	7	2	14
14 - 15	14,5	6	3	18
15 - 16	15,5	1	4	4
16 - 17	-16,5	3	5	15
Tổng		42		43

$$\bar{x} = 11,5 + 1 \left(\frac{43}{42} \right) \approx 12,524$$

Trị trung bình có hai đặc điểm đáng chú ý là tổng độ lệch từ các số liệu tới trị trung bình bằng zero, tức $\sum_{i=1}^n (x_i - \bar{x}) = 0$ và $\sum_{i=1}^n (x_i - \bar{x})^2 = \min$, tức $\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - A)^2$ với A là một trị số bất kỳ nào đó, dấu “=” chỉ xảy ra khi và chỉ khi $A = \bar{x}$.

Phụ chú 1.1: Ngoài trung bình cộng ra còn có nhiều loại trung bình khác nữa như trung bình nhân, trung bình điều hòa, trung bình bình phương, trung bình căn bậc hai... cũng có thể được dùng làm đại diện cho mẫu hoặc quần thể, đặc biệt là đối với các quần thể không phải là kết quả của sự thăng giáng ngẫu nhiên. Tương tự như trung bình cộng, các trung bình khác cũng được định nghĩa dựa trên nguyên lý chung (nguyên lý moment) như sau:

* *Trung bình nhân:*

$$\underbrace{\underbrace{x_1 \cdot x_1 \dots x_1}_{n_1} \cdot \underbrace{x_2 \cdot x_2 \dots x_2}_{n_2} \cdot \underbrace{x_K \cdot x_K \dots x_K}_{n_K}}_n = \bar{x} \cdot \bar{x} \dots \bar{x}_n$$

hay $\prod_{i=1}^K x_i^{n_i} = \bar{x}^n$ hay $\bar{x} = \sqrt[n]{\prod_{i=1}^K x_i^{n_i}}$, với $n = \sum_{i=1}^K n_i$

* Trung bình điều hòa:

$$\underbrace{\frac{1}{x_1} + \frac{1}{x_1} + \dots + \frac{1}{x_1}}_{n_1} + \underbrace{\frac{1}{x_2} + \frac{1}{x_2} + \dots + \frac{1}{x_2}}_{n_2} + \underbrace{\frac{1}{x_K} + \frac{1}{x_K} + \dots + \frac{1}{x_K}}_{n_K} =$$

$$\underbrace{\hspace{10em}}_n$$

$$= \underbrace{\frac{1}{\bar{x}} + \frac{1}{\bar{x}} + \dots + \frac{1}{\bar{x}}}_n \text{ hay } \sum_{i=1}^K \frac{1}{x_i} = n \cdot \frac{1}{\bar{x}} \text{ hay } \bar{x} = \frac{n}{\sum_{i=1}^K \left(\frac{1}{x_i}\right)}, \text{ với } n = \sum_{i=1}^K n_i$$

* Trung bình bình phương:

$$\underbrace{x_1^2 + x_1^2 + \dots + x_1^2}_{n_1} + \underbrace{x_2^2 + x_2^2 + \dots + x_2^2}_{n_2} + \underbrace{x_K^2 + x_K^2 + \dots + x_K^2}_{n_K} =$$

$$\underbrace{\hspace{10em}}_n$$

$$= \underbrace{\bar{x}^2 + \bar{x}^2 + \dots + \bar{x}^2}_n$$

hay $\sum_{i=1}^K n_i x_i^2 = n \bar{x}^2$ hay $\bar{x} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^K n_i x_i^2}$, với $n = \sum_{i=1}^K n_i$

* Trung bình căn bậc hai:

$$\underbrace{\sqrt{x_1} + \sqrt{x_1} + \dots + \sqrt{x_1}}_{n_1} + \underbrace{\sqrt{x_2} + \sqrt{x_2} + \dots + \sqrt{x_2}}_{n_2} + \underbrace{\sqrt{x_K} + \sqrt{x_K} + \dots + \sqrt{x_K}}_{n_K} =$$

$$\underbrace{\hspace{10em}}_n$$

$$= \underbrace{\sqrt{\bar{x}} + \sqrt{\bar{x}} + \dots + \sqrt{\bar{x}}}_n$$

hay $\sum_{i=1}^K n_i \sqrt{x_i} = n \sqrt{\bar{x}}$ hay $\bar{x} = \left[\frac{1}{n} \sum_{i=1}^K n_i \cdot \sqrt{x_i} \right]^2$, với $n = \sum_{i=1}^K n_i$

Để tạo thuận lợi cho việc trình bày các phần tiếp theo chúng ta cần xem xét tới vấn đề thế nào là một ước lượng và thế nào là một ước lượng tốt nhất đặc biệt là các ước lượng và ước lượng tốt nhất của trung bình (mean) và biến lượng (variance).

Như chúng ta biết, hàm mật độ của biến ngẫu nhiên X của phân phối bình thường dạng tổng quát $N(m, \sigma)$ là:

$$f_x(x, m, \sigma^2) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x_i - m)^2}{2\sigma^2}}$$

với m là trung bình và σ^2 là biến lượng của phân phối. Hàm “chắc đúng” (likelihood) $L(m, \sigma^2)$ được dùng để xác định ước lượng của trung bình ($\bar{m}$) và ước lượng của biến lượng ($\hat{\sigma}^2$) sẽ như sau:

$$L(m, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x_i - m)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{\sum (x_i - m)^2}{2\sigma^2}}$$

$$\text{hay} \quad \ln L(m, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{\sum (x_i - m)^2}{2\sigma^2}$$

Từ điều kiện cực đại của hàm $\ln L(m, \sigma^2)$, chúng ta dễ dàng xác định được các ước lượng chắc đúng nhất của trung bình và biến lượng thường được gọi là phương pháp ước lượng “chắc đúng nhất” (maximum likelihood) như sau:

$$\begin{aligned} \frac{\partial \ln L(m, \sigma^2)}{\partial m} &= \frac{1}{\sigma^2} \sum (x_i - m) = 0 \\ \frac{\partial \ln L(m, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (x_i - m)^2 = 0 \end{aligned}$$

Từ phương trình thứ nhất của hệ ta suy ra được ước lượng của trung bình là $\bar{m} = \frac{1}{n} \sum x_i = \bar{x}$. Thế trị này vào phương trình thứ hai của hệ ta sẽ có được ước lượng của biến lượng là $\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = S^{*2}$. Bây giờ chúng ta tiếp tục xét xem liệu đây có phải là các ước lượng tốt nhất hay không.

$$\text{Vì rằng:} \quad M[\bar{x}] = M\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n M[x_i] = \frac{1}{n} \cdot n m = m$$

$$\text{và} \quad V[\bar{x}] = V\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n^2} \sum_{i=1}^n V[x_i] = \frac{1}{n^2} \cdot n \sigma^2 = \frac{\sigma^2}{n}$$

Do đó, khi $n \rightarrow \infty$ thì $V[\bar{x}] \rightarrow 0$, điều này có nghĩa rằng trung bình của mẫu tính theo công thức $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ là ước lượng tốt nhất cho trung bình của quần thể m .

$$\begin{aligned}
\text{Vì rằng: } \sum_{i=1}^n (x_i - m)^2 &= \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - m)]^2 = \\
&= \sum_{i=1}^n [(x_i - \bar{x})^2 + (\bar{x} - m)^2 + 2(\bar{x} - m)(x_i - \bar{x})] \\
&= \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (\bar{x} - m)^2 + 2(\bar{x} - m) \sum_{i=1}^n (x_i - \bar{x}) \\
&= \sum_{i=1}^n (x_i - \bar{x})^2 + n \sum_{i=1}^n (\bar{x} - m)^2 \quad (\text{vì } \sum_{i=1}^n (x_i - \bar{x}) = 0)
\end{aligned}$$

Từ đó suy ra được hệ thức sau đây:

$$\underbrace{\frac{1}{n} \sum_{i=1}^n (x_i - m)^2}_{\sigma^2} = \underbrace{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 + (\bar{x} - m)^2}_{S^{*2}}$$

Điều này có nghĩa là $S^{*2} > \sigma^2$, như vậy rõ ràng là biến lượng của mẫu tính theo công thức $S^{*2} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ không phải là ước lượng tốt nhất cho biến lượng của quần thể σ^2 .

Mặt khác, F. W. Bessel đã chứng minh được rằng ước lượng tốt nhất của biến lượng là $\tilde{\sigma}^2 = \frac{n}{n-1} S^{*2}$ và lượng $\frac{n}{n-1}$ được gọi là tu chỉnh Bessel (Bessel's correction).

Vì rằng: $\tilde{\sigma}^2 = \frac{n}{n-1} S^{*2} = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = S^2$,
 $M[S^2] = \sigma^2$ và $V[S^2] = \frac{2}{n-1} \sigma^4$ do đó, khi $n \rightarrow \infty$ thì $V[S^2] \rightarrow 0$, điều này một lần nữa chứng tỏ rằng biến lượng của mẫu tính theo công thức $S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ là ước lượng tốt nhất cho biến lượng của quần thể σ^2 và lượng $n-1$ được R. A. Fisher gọi là “độ tự do” (degrees of freedom).

Vì rằng $S^2 = \frac{n}{n-1} S^{*2}$, do đó, trong trường hợp mẫu lớn ($n > 30$) chúng ta có thể coi $S^2 \approx S^{*2}$.

1.2. BIẾN LƯỢNG VÀ ĐỘ LỆCH CHUẨN

Nếu x_i là số liệu của các thăng giáng ngẫu nhiên được lấy ra từ một quần thể nào đó và n_i là số số liệu giống nhau của mỗi trị số x_i , còn n là tổng số các số liệu được lấy ra thì biến lượng của mẫu, ký hiệu là S^2 , được định nghĩa như sau:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^K n_i (x_i - \bar{x})^2 \quad \text{với } n = \sum_{i=1}^K n_i$$

Ở đây $n-1$ là độ tự do, còn S là độ lệch chuẩn (Standard deviation) của mẫu.

Bây giờ nếu như kích thước của mẫu bằng kích thước của quần thể, hay $n \approx \infty$, thì trung bình của mẫu ($\bar{x}$) trở thành trung bình của quần thể (m), biến lượng của mẫu (S^2) trở thành biến lượng của quần thể (σ^2) và độ lệch chuẩn của mẫu (S) trở thành độ lệch chuẩn của quần thể (σ).

Để giúp cho việc tính toán biến lượng của mẫu được dễ dàng hơn chúng ta cần chú ý tới công thức tính sau đây:

$$\begin{aligned} \sum_{i=1}^K n_i (x_i - \bar{x})^2 &= \sum_{i=1}^K (n_i x_i^2 + n_i \bar{x}^2 - 2n_i x_i \bar{x}) \\ &= \sum_{i=1}^K n_i x_i^2 + \bar{x}^2 \sum_{i=1}^K n_i - 2\bar{x} \sum_{i=1}^K n_i x_i \\ &= \sum_{i=1}^K n_i x_i^2 + n \left(\frac{1}{n} \sum_{i=1}^K n_i x_i \right)^2 - \frac{2}{n} \left(\sum_{i=1}^K n_i x_i \right) \left(\sum_{i=1}^K n_i x_i \right) \\ &= \sum_{i=1}^K n_i x_i^2 - \frac{1}{n} \left(\sum_{i=1}^K n_i x_i \right)^2 \end{aligned}$$

Trong nhiều trường hợp để đơn giản hóa công việc tính toán, chúng ta thường tiến hành phân nhóm và mã hóa số liệu, do đó, chúng ta cần nắm vững kỹ thuật tính toán để tránh những sai sót đáng tiếc có thể xảy ra.

Giả sử, chúng ta mã hóa số liệu theo công thức tổng quát là $u_i = \alpha(x_i - x_0)$ hay $x_i = x_0 + \frac{1}{\alpha} u_i$, do đó, $V[x_i] = V\left[x_0 + \frac{1}{\alpha} u_i\right] = \frac{1}{\alpha^2} V[u_i]$ hay $S_x^2 = \frac{1}{\alpha^2} S_u^2$.

Như đã nói ở trên, chúng ta thường chọn $\alpha = \frac{1}{b}$ với b là độ lớn của nhóm.

Chú ý rằng, biến lượng của mẫu được tính dựa trên số liệu phân nhóm

thường lớn hơn một chút so với biến lượng của mẫu được tính trực tiếp từ số liệu thu được, do đó W. F. Sheppard đã đề nghị một lượng giảm là $(-\frac{b^2}{12})$, với b là độ lớn của nhóm, và lượng giảm này được gọi là lượng tu chỉnh Sheppard (Sheppard's correction). Dù vậy, R. A. Fisher vẫn cho rằng chúng ta không nên sử dụng lượng tu chỉnh Sheppard $(-\frac{b^2}{12})$ để tính biến lượng của mẫu bởi lẽ vấn đề ở đây không những chỉ liên quan tới biến lượng của mẫu mà còn liên quan tới trung bình của mẫu. Do đó, nếu như chúng ta sử dụng một biến lượng đã được tu chỉnh đồng thời với một trung bình không được tu chỉnh thì rất có khả năng dẫn tới những kết luận sai lầm đáng tiếc trong các trắc nghiệm thống kê học. Số liệu ở bảng 1.2 và kết quả tính toán sau đây sẽ minh họa cho những điều trình bày ở trên.

Vì rằng: $n = 42, \sum x_i = 522,23, \sum x_i^2 = 6652,25$

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 6652,25 - \frac{(522,23)^2}{42} \approx 158,8173$$

do đó $S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{1}{42-1} \cdot 158,8173 \approx 3,8736$.

Bây giờ chúng ta tiến hành tính toán theo số liệu đã phân nhóm như sau:

Vì rằng: $n = 42, \sum n_i u_i = 43, \sum n_i u_i^2 = 215,$

$$\sum n_i (u_i - \bar{u})^2 = \sum n_i u_i^2 - \frac{(\sum n_i u_i)^2}{n} = 215 - \frac{(43)^2}{42} \approx 170,976$$

Do đó, $S^2 = b^2 S_u^2 = b^2 \cdot \frac{1}{n-1} \sum n_i (u_i - \bar{u})^2 = 1^2 \cdot \frac{1}{42-1} \cdot 170,976 \approx 4,1701$

Ở đây, nếu chúng ta sử dụng lượng tu chỉnh Sheppard thì biến lượng của mẫu sẽ là $4,1701 - \frac{1^2}{12} \approx 4,0868$. Kết quả được tóm tắt ở bảng sau đây:

	Không phân nhóm số liệu	Có phân nhóm số liệu
Trung bình	12,434	12,524
Biến lượng	3,8736	4,1701 (Không tu chỉnh)
		4,0868 (Có tu chỉnh)

1.3. GIAO BIẾN LƯỢNG (COVARIANCE)

Nếu (x_i, y_i) là cặp số liệu được lấy ra từ quần thể nhị biến (bivariate population) và n là tổng số các cặp số liệu được lấy ra thì *giao biến lượng* của mẫu, ký hiệu là K_{xy} , được định nghĩa như sau:

$$K_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Vì rằng:

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n (x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x} \bar{y}) = \\ &= \sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i - \bar{y} \sum_{i=1}^n x_i + n \bar{x} \bar{y} = \\ &= \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) - \frac{1}{n} \left(\sum_{i=1}^n y_i \right) \left(\sum_{i=1}^n x_i \right) + \\ &\quad + n \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \frac{1}{n} \left(\sum_{i=1}^n y_i \right) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \end{aligned}$$

$$\text{Do đó: } K_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \right]$$

Bây giờ chúng ta xét tới trường hợp số liệu có phân nhóm. Giả sử số liệu được phân thành $K \times I$ nhóm và x_i ($i = 1, 2, \dots, k$) và y_j ($j = 1, 2, \dots, I$) là các trị trung tâm của các phân nhóm, còn n_{ij} là tần số của phân nhóm (x_i, y_j) , khi đó $\sum_{i=1}^K \sum_{j=1}^I n_{ij} = n$. Với $\sum_{i=1}^K n_{ij} = n_j$, $\sum_{j=1}^I n_{ij} = n_i$, và số liệu được mã hóa theo

công thức $u_i = \frac{x_i - x_0}{b_x}$, ($i = 1, 2, \dots, k$) và $v_j = \frac{y_j - y_0}{b_y}$, ($j = 1, 2, \dots, I$) với x_0

và y_0 thường được chọn tại vị trí có tần số n_{ij} lớn nhất, tương tự như trên chúng ta dễ dàng chứng minh được là:

$$K_{uv} = \frac{1}{n-1} \sum_{i=1}^K \sum_{j=1}^I n_{ij} (u_i - \bar{u})(v_j - \bar{v})$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^K \sum_{j=1}^l n_{ij} u_i v_j - \frac{\left(\sum_{i=1}^K n_i u_i \right) \left(\sum_{j=1}^l n_j v_j \right)}{n} \right]$$

Vì rằng: $K_{xy} = K[x_i y_i] = K[(x_o + b_x u_i)(y_o + b_y v_j)] = K[b_x u_i b_y v_j] = b_x b_y K[u_i v_j] = b_x b_y K_{uv}$, do đó, chúng ta dễ dàng tính được giao biến lượng K_{xy} thông qua việc tính giao biến lượng K_{uv} .

Chú ý rằng nếu như các biến ngẫu nhiên X và Y độc lập với nhau, nghĩa là mối quan hệ nhân quả $X \leftrightarrow Y$ không tồn tại, thì giao biến lượng $K_{xy} = 0$.

$$\text{Vì rằng: } K_{xx} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = S_x^2$$

Do đó: $V[x] = K[xx]$.

Công thức này giúp chúng ta có một cái nhìn toàn diện hơn về vấn đề biến lượng và giao biến lượng. Các ví dụ minh họa về giao biến lượng sẽ được trình bày ở phần tương quan và hồi quy.

Trước khi kết thúc chương này tôi xin tóm tắt một số điểm quan trọng, đáng nhớ sau đây:

1 - Các đại lượng thống kê

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$K_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

2 - Các công thức tính toán

$$Q_x = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2$$

$$Q_y = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2$$

$$Q_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)$$

3 - Các kỹ thuật tính toán

$$M \left[\sum_{K=1}^n a_K x_K + b \right] = \sum_{K=1}^n a_K M[x_K] + b$$

Ví dụ: $M[ax + by + c] = aM[x] + bM[y] + c$

hay $\overline{ax + by + c} = a\bar{x} + b\bar{y} + c$

$$V \left[\sum_{K=1}^n a_K x_K + b \right] = \sum_{K=1}^n a_K^2 V[x_K] + 2 \sum_{\substack{i,j=1 \\ i < j}}^n a_i a_j K_{ij}$$

Ví dụ: $V[ax + by + c] = a^2 V[x] + b^2 V[y] + 2abK_{xy}$

hay $S_{ax+by+c}^2 = a^2 S_x^2 + b^2 S_y^2 + 2abK_{xy}$

Thực ra, chúng ta có thể dùng thuật tính giao biến lượng để tính biến lượng như sau:

$$\begin{aligned} V[ax + by + c] &= K[(ax + by + c)(ax + by + c)] \\ &= K[a^2 x^2 + b^2 y^2 + 2abxy + 2acx + 2bcy + c^2] \\ &= a^2 K[x^2] + b^2 K[y^2] + 2abK[xy] \\ &= a^2 V[x] + b^2 V[y] + 2abK_{xy}. \end{aligned}$$

Chương 2

CÁC PHÂN PHỐI TRONG THỐNG KÊ HỌC

2.1. PHÂN PHỐI NHỊ THỨC, PHÂN PHỐI POISSON VÀ PHÂN PHỐI BÌNH THƯỜNG

2.1.1. Đặc điểm

Giả sử rằng xác suất xuất hiện của biến ngẫu nhiên X trong mỗi phép thử là p thì xác suất để trong n phép thử biến ngẫu nhiên X xuất hiện $0, 1, 2, \dots, x, \dots, n$ lần được xác định bởi công thức Bernoulli như sau:

$$f(x) = P\{X = x\} = C_n^x p^x (1-p)^{n-x} = C_n^x p^x q^{n-x}, \text{ với } q = 1 - p$$

Luật phân phối xác suất của biến ngẫu nhiên X được mô tả bởi công thức Bernoulli được gọi là luật phân phối nhị thức hay phân phối Bernoulli. Trung bình và biến lượng của phân phối này được xác định như sau:

x	0	1	2	...	$n-1$	n	Tổng
$f(x) = C_n^x p^x q^{n-x}$	q^n	npq^{n-1}	$\frac{n(n-1)}{12} p^2 q^{n-2}$	...	$np^{n-1}q$	p^n	$(p+q)^n = 1$
$xf(x)$	0	npq^{n-1}	$n(n-1)p^2 q^{n-2}$	...	$n(n-1)p^{n-1}q$	np^n	$np(p+q)^{n-1} = np$
$x(x-1)f(x)$	0	0	$n(n-1)p^2 q^{n-2}$	...	$n(n-1)(n-2)p^{n-1}q$	$n(n-1)p^n$	

$$\sum f(x) = q^n + npq^{n-1} + \frac{n(n-1)}{12} p^2 q^{n-2} + \dots + p^n = (p+q)^n = 1$$

$$\sum xf(x) = np[q^{n-1} + (n-1)pq^{n-2} + \frac{(n-1)(n-2)}{1 \cdot 2} p^2 q^{n-3} + \dots + p^{n-1}] =$$

$$= np(p+q)^{n-1} = np$$

$$\begin{aligned} \sum x(x-1)f(x) &= n(n-1)p^2[q^{n-2} + (n-2)pq^{n-3} + \dots + p^{n-2}] = \\ &= n(n-1)p^2(p+q)^{n-2} = n(n-1)p^2 \end{aligned}$$

$$\begin{aligned}\Sigma x^2 f(x) &= \Sigma [x(x-1) + x] f(x) = \Sigma x(x-1) f(x) + \Sigma x f(x) = \\ &= n(n-1)p^2 + np = n^2 p^2 + npq\end{aligned}$$

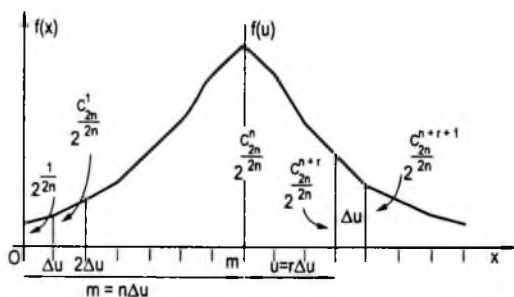
$$m = \frac{\Sigma x f(x)}{\Sigma f(x)} = \frac{np}{1} = np$$

$$\begin{aligned}\sigma^2 &= \frac{\Sigma (x-m)^2 f(x)}{\Sigma f(x)} = \frac{\Sigma (x^2 - 2mx + m^2) f(x)}{\Sigma f(x)} = \\ &= \frac{\Sigma x^2 f(x)}{\Sigma f(x)} - \frac{2m \Sigma x f(x)}{\Sigma f(x)} + \frac{m^2 \Sigma f(x)}{\Sigma f(x)} = \frac{n^2 p^2 + npq}{1} - 2m \cdot m + m^2 = \\ &= n^2 p^2 + npq - m^2 = n^2 p^2 + npq - (np)^2 = npq\end{aligned}$$

Để thấy được rằng *phân phối bình thường (phân phối Gauss)* là giới hạn của phân phối nhị thức khi số phép thử tăng lên vô hạn chúng ta tiến hành triển khai phân phối nhị thức dạng $B(2n, \frac{1}{2})$ rồi phác họa trên đồ thị (hình

2.1) như sau:

$$\left(\frac{1}{2} + \frac{1}{2}\right)^{2n} = \frac{1}{2^{2n}} [1 + C_{2n}^1 + C_{2n}^2 + \dots + C_{2n}^n + \dots + C_{2n}^{n+r} + \dots + 1]$$



Hình 2.1. Đồ thị phác họa phân phối nhị thức dạng $B(2n, \frac{1}{2})$

Tiến hành tính toán tương tự như trên chúng ta sẽ xác định được trung bình và biến lượng của phân phối nhị thức dạng $B(2n, \frac{1}{2})$ này như sau:

$$m = n \Delta u; \sigma^2 = \frac{n}{2} (\Delta u)^2$$

Vì rằng:

$$f(u) = \frac{1}{2^{2n}} C_{2n}^{n+r}$$

$$f(u+\Delta u) = \frac{1}{2^{2n}} C_{2n}^{n+r+1} = \frac{1}{2^{2n}} \cdot \frac{n-r}{n+r+1} C_{2n}^{n+r}$$

do đó: $\frac{\Delta f(u)}{\Delta u} = \left(\frac{-2r-1}{n+r+1} \right) \cdot \frac{f(u)}{\Delta u}$. Chú ý là $u = r\Delta u \Leftrightarrow r = \frac{u}{\Delta u}$

và $\sigma^2 = \frac{n}{2}(\Delta u)^2 \Leftrightarrow n(\Delta u)^2 = 2\sigma^2$

do đó: $\frac{\Delta f(u)}{\Delta u} = \left(\frac{-2r-1}{n+r+1} \right) \cdot \frac{f(u)}{\Delta u} = -f(u) \left[\frac{2u + \Delta u}{2\sigma^2 + u\Delta u + (\Delta u)^2} \right]$

Vì rằng khi $n \rightarrow \infty$ thì $\Delta u \rightarrow 0$ nên ta có được đẳng thức sau đây:

$$\frac{\partial f(u)}{\partial u} = -f(u) \cdot \frac{u}{\sigma^2} \text{ hay } \frac{\partial f(u)}{\partial u} = \frac{-1}{\sigma^2} \cdot u \partial u$$

Từ đó suy ra: $f(u) = C e^{\frac{-u^2}{2\sigma^2}}$

Chú ý là $\int_{-\infty}^{+\infty} f(u) du = 1$ và $\int_{-\infty}^{+\infty} e^{-u^2} du = \sqrt{\pi}$, nên $C = \frac{1}{\sigma\sqrt{2\pi}}$

Do đó: $f(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma^2}}$

Chú ý rằng u là độ lệch từ trung bình m , tức $u = x - m$,

Do đó: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$ với $-\infty < x < +\infty$

Đây chính là hàm mật độ của phân phối bình thường dạng tổng quát $N(m, \sigma)$ đã được đề cập ở chương 1. Trong trường hợp $m = 0$ và $\sigma = 1$ chúng ta sẽ có được hàm mật độ của phân phối bình thường dạng chuẩn $N(0, 1)$ là:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Từ đó chúng ta xác định được hàm *phân số xác suất* của phân phối bình thường như sau:

$$\Phi(x) = \int_{-\infty}^x f(t) dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$

Thực ra chúng ta cũng có thể xác định được hàm phân phối xác suất của phân phối bình thường từ dạng tổng quát $N(m, \sigma)$ như sau:

với $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$, ta đặt $\frac{x-m}{\sigma} = t$, suy ra: $\frac{\partial x}{\sigma} = \partial t$, do đó:

$$\Phi(x) = \int_{-\infty}^x f(x) \partial x = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(x-m)^2}{2\sigma^2}} \partial x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} \partial t$$

Bây giờ chúng ta tiến hành xem xét mối quan hệ giữa phân phối nhị thức và *phân phối Poisson*.

$$\begin{aligned} \text{Vì rằng: } f(x) = P\{X = x\} &= C_n^x p^x q^{n-x} = \frac{n!}{x!(n-x)!} \cdot p^x q^{n-x} = \\ &= \frac{n(n-1)(n-2)\dots(n-x+1)}{x!} \cdot p^x (1-p)^{n-x} = \\ &= \frac{(np)^x \cdot \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{x-1}{n}\right)}{x!} \cdot (1-p)^p \end{aligned}$$

nên khi $n \rightarrow \infty$, $p \rightarrow 0$, nhưng $np = \lambda$ thì $f(x) = P\{X = x\} = \frac{\lambda^x}{x!} e^{-\lambda}$

Luật phân phối xác suất của biến ngẫu nhiên X được mô tả bởi công thức trên được gọi là luật phân phối Poisson. Tiến hành tính toán như đã được trình bày đối với phân phối nhị thức chúng ta sẽ xác định được *trung bình và biến lượng của phân phối Poisson đều bằng λ* .

Để có thể xử lý số liệu thống kê được dễ dàng chúng ta cần nắm vững kỹ thuật tính toán để xác định trung bình và biến lượng của các phân phối xác suất trong các trường hợp thông thường sau đây:

1 - Nếu biến ngẫu nhiên X có phân phối nhị thức dạng $B(n, p)$ với trung bình $M[x] = np$ và biến lượng $V[x] = npq$ thì:

- Đại lượng x/n cũng có phân phối nhị thức với trung bình và biến lượng như sau:

$$M\left[\frac{x}{n}\right] = \left[\frac{1}{n}\right] M[x] = \frac{1}{n} \cdot np = p$$

$$V\left[\frac{x}{n}\right] = \frac{1}{n^2} V[x] = \frac{1}{n^2} \cdot npq = \frac{pq}{n}$$

- Đại lượng $\frac{\frac{x}{n} - p}{\sqrt{\frac{pq}{n}}}$ cũng có phân phối nhị thức với trung bình và

biến lượng như sau:

$$M\left[\frac{\frac{x}{n} - p}{\sqrt{\frac{pq}{n}}}\right] = \frac{1}{\sqrt{\frac{pq}{n}}} \cdot M\left[\frac{x}{n} - p\right] = \frac{1}{\sqrt{\frac{pq}{n}}} \left(M\left[\frac{x}{n}\right] - M[p] \right) = \frac{1}{\sqrt{\frac{pq}{n}}} (p - p) = 0$$

$$V\left[\frac{\frac{x}{n} - p}{\sqrt{\frac{pq}{n}}}\right] = \frac{1}{\left(\sqrt{\frac{pq}{n}}\right)^2} \cdot V\left[\frac{x}{n} - p\right] = \frac{1}{\frac{pq}{n}} \cdot V\left[\frac{x}{n}\right] = \frac{1}{\frac{pq}{n}} \cdot \frac{pq}{n} = 1$$

Tương tự, chúng ta có thể xác định trung bình và biến lượng của đại lượng $\frac{x - np}{\sqrt{npq}}$.

2 - Nếu X_1 và X_2 là các biến ngẫu nhiên *độc lập* có phân phối nhị thức dạng $B(n_1, p)$ và $B(n_2, p)$ thì:

- Đại lượng $\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)$ cũng có phân phối nhị thức với trung bình và biến

lượng như sau:

$$M\left[\frac{x_1}{n_1} - \frac{x_2}{n_2}\right] = M\left[\frac{x_1}{n_1}\right] - M\left[\frac{x_2}{n_2}\right] = p - p = 0$$

$$V\left[\frac{x_1}{n_1} - \frac{x_2}{n_2}\right] = V\left[\frac{x_1}{n_1}\right] + V\left[\frac{x_2}{n_2}\right] = \frac{pq}{n_1} + \frac{pq}{n_2} = pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

- Đại lượng $\frac{\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ cũng có phân phối nhị thức với trung bình và

biến lượng như sau:

$$M\left[\frac{\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\right] = \frac{1}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \cdot M\left[\frac{x_1}{n_1} - \frac{x_2}{n_2}\right] =$$

$$= \frac{1}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \cdot 0 = 0$$

$$V\left[\frac{\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\right] = \frac{1}{\left(\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)^2} \cdot V\left[\frac{x_1}{n_1} - \frac{x_2}{n_2}\right] =$$

$$= \frac{1}{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \cdot pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 1$$

Như đã trình bày ở trên, phân phối bình thường là giới hạn của phân phối nhị thức khi $n \rightarrow \infty$, do đó, trong trường hợp n lớn hay $np, np \geq 5$ thì:

- x sẽ có phân phối cận bình thường dạng $N(np, \sqrt{npq})$.

- $\frac{x}{n}$ sẽ có phân phối cận bình thường dạng $N\left(p, \sqrt{\frac{pq}{n}}\right)$.

- $\frac{\left(\frac{x}{n} - p\right)}{\sqrt{\frac{pq}{n}}}$ hay $\frac{(x - np)}{\sqrt{npq}}$ sẽ có phân phối cận bình thường dạng $N(0, 1)$.

- $\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)$ sẽ có dạng phân phối cận bình thường dạng:

$$N\left(0, \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

- $\frac{\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ sẽ có phân phối cận bình thường dạng $N(0, 1)$.

3 - Nếu biến ngẫu nhiên X có phân phối bình thường dạng $N(m, \sigma)$ với trung bình $M[x] = m$ và biến lượng $V[x] = \sigma^2$ thì:

- Đại lượng $\frac{(x-m)}{\sigma}$ cũng có phân phối bình thường với trung bình và biến lượng như sau:

$$M\left[\frac{x-m}{\sigma}\right] = \frac{1}{\sigma} M[x-m] = \frac{1}{\sigma} (M[x] - M[m]) = \frac{1}{\sigma} (m-m) = 0$$

$$V\left[\frac{x-m}{\sigma}\right] = \frac{1}{\sigma^2} V[x-m] = \frac{1}{\sigma^2} V[x] = \frac{1}{\sigma^2} \cdot \sigma^2 = 1$$

- Đại lượng $\bar{x}$ cũng có phân phối bình thường với trung bình và biến lượng như sau:

$$M[\bar{x}] = M\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n M[x_i] = \frac{1}{n} \cdot nm = m$$

$$V[\bar{x}] = M\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n^2} \sum_{i=1}^n V[x_i] = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

- Đại lượng $\frac{(\bar{x}-m)}{\sqrt{\frac{\sigma^2}{n}}}$ cũng có phân phối bình thường với trung bình và

biến lượng như sau:

$$M\left[\frac{\bar{x}-m}{\sqrt{\frac{\sigma^2}{n}}}\right] = \frac{1}{\sqrt{\frac{\sigma^2}{n}}} M[\bar{x}-m] = \frac{1}{\sqrt{\frac{\sigma^2}{n}}} (M[\bar{x}] - M[m]) = \frac{1}{\sqrt{\frac{\sigma^2}{n}}} (m-m) = 0$$

$$V\left[\frac{\bar{x}-m}{\sqrt{\frac{\sigma^2}{n}}}\right] = \frac{1}{\left(\sqrt{\frac{\sigma^2}{n}}\right)^2} V[\bar{x}-m] = \frac{1}{\frac{\sigma^2}{n}} \cdot V[\bar{x}] = \frac{1}{\frac{\sigma^2}{n}} \cdot \frac{\sigma^2}{n} = 1$$

4 - Nếu X_1 và X_2 là các biến ngẫu nhiên độc lập có phân phối bình thường dạng $N(m_1, \sigma)$ và $N(m_2, \sigma)$ thì:

- Đại lượng $(x_1 - x_2)$ cũng có phân phối bình thường với trung bình và biến lượng như sau:

$$M[x_1 - x_2] = M[x_1] - M[x_2] = m_1 - m_2$$

$$V[x_1 - x_2] = V[x_1] + V[x_2] = \sigma^2 + \sigma^2 = 2\sigma^2$$

- Đại lượng $(\bar{x}_1 - \bar{x}_2)$ cũng có phân phối bình thường với trung bình và biến lượng như sau:

$$M[\bar{x}_1 - \bar{x}_2] = M[\bar{x}_1] - M[\bar{x}_2] = m_1 - m_2$$

$$V[\bar{x}_1 - \bar{x}_2] = V[\bar{x}_1] + V[\bar{x}_2] = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

- Đại lượng $\frac{[(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)]}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ cũng có phân phối bình thường với

trung bình biến lượng như sau:

$$M\left[\frac{[(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)]}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}\right] = \frac{1}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} M[(\bar{x}_1 - \bar{x}_2) - m_1 + m_2]$$

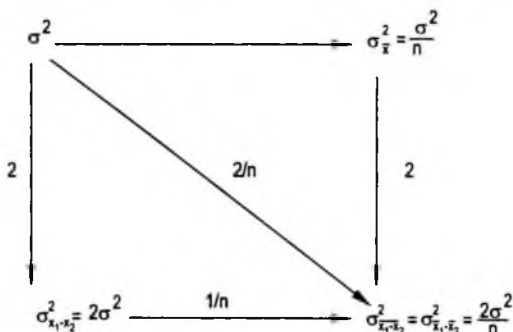
$$= \frac{1}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} (M[\bar{x}_1 - \bar{x}_2] - M[m_1] + M[m_2])$$

$$= \frac{1}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} (m_1 - m_2 - m_1 + m_2) = 0$$

$$V \left[\frac{[(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)]}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \right] = \frac{1}{\left(\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right)^2} V[(\bar{x}_1 - \bar{x}_2) - m_1 + m_2]$$

$$= \frac{1}{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \cdot V[\bar{x}_1 - \bar{x}_2] = \frac{1}{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \cdot \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = 1$$

Mối quan hệ giữa các biến lượng của phân phối bình thường dạng $N(m, \sigma)$, $N(m_1, \sigma)$ và $N(m_2, \sigma)$, với $n_1 = n_2 = n$, có thể được tóm tắt bởi đồ sau đây:

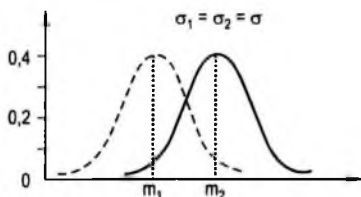


2.1.2. Ứng dụng

2.1.2.1. Trắc nghiệm U

Phân phối bình thường và cận bình thường đóng một vai trò rất quan trọng trong thống kê học. Trắc nghiệm so sánh dựa trên căn bản của phân phối bình thường và cận bình thường dạng $N(0, 1)$ được gọi là “trắc nghiệm U”.

Chú ý rằng, hầu hết các trắc nghiệm so sánh đều dựa trên căn bản là biến lượng của quần thể thì bằng nhau, tức $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_l^2 (= \sigma^2)$, còn trung bình là đại lượng đem so sánh thì có thể bằng nhau hoặc khác nhau. Do đó, trong thuật so sánh chúng ta chỉ cần đặt giả thuyết “*tương đồng H_0* ” rằng các trung bình bằng nhau, tức $H_0 : m_1 = m_2 = \dots = m_l (= m)$ và tiến hành kiểm tra xem giả thuyết H_0 này là đúng hay sai bằng cách dựa vào đặc điểm của các phân phối thường gặp trong thống kê học. Nhờ đó chúng ta mới có cơ sở để *chấp nhận* hoặc *bác bỏ* giả thuyết H_0 (hình 2.2).

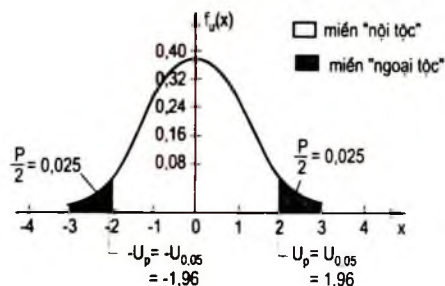


Hình 2.2. Đồ thị minh họa giả thuyết tương đồng $H_0 : m_1 = m_2$

Đặc điểm của phân phối mà chúng ta thường dựa vào đó kiểm tra giả thuyết H_0 là đúng hay sai nhằm có một kết luận xác đáng là nên chấp nhận hoặc là phải bác bỏ giả thuyết H_0 này là trung bình của phân phối, đặc biệt là *các phân phối có trung bình là 0* như phân phối bình thường và cận bình thường dạng chuẩn $N(0, 1)$, phân phối Student $t(K)$ và *các phân phối có trung bình là 1* hay *xấp xỉ 1* như phân phối Fisher $F(k_1, k_2)$ và phân phối $\frac{\chi^2(K)}{K}$.

Chú ý rằng, hầu hết các trắc nghiệm so sánh trong thống kê học đều tiến hành theo phương cách hoặc là *thiết lập hiệu số* hoặc là *thiết lập tỷ số* giữa các đại lượng cần so sánh (xem chương 3). Trong trường hợp các trắc nghiệm so sánh được thực hiện bằng cách thiết lập hiệu số thì các phân phối có trung bình là zero được áp dụng để kiểm tra giả thuyết H_0 , còn trong trường hợp các trắc nghiệm so sánh được thiết lập bằng cách thiết lập tỷ số thì các phân phối có trung bình là 1 hay xấp xỉ 1 được áp dụng để kiểm tra giả thuyết H_0 . Như vậy, để có thể tiến hành so sánh, trước tiên chúng ta phải thiết lập được một biểu thức trong đó các đại lượng cần so sánh nằm dưới dạng hiệu số hoặc tỷ số, sau đó là xác định quy luật phân phối của nó và căn cứ vào đặc điểm của trung bình của phân phối đó để chấp nhận hoặc bác bỏ giả thuyết H_0 . Trong trường hợp so sánh được thực hiện bằng cách thiết lập hiệu số giữa các đại lượng cần so sánh, nếu như hiệu của hai đại lượng cần

so sánh nằm ở *miền lân cận giá trị trung bình (bằng zero) của phân phối (miền ổn định chấp nhận giả thuyết tương đồng H_0)* thì hai đại lượng đem so sánh có thể coi là như nhau. Trái lại, nếu hiệu của hai đại lượng cần so sánh nằm ngoài miền này (miền ổn định bác bỏ giả thuyết tương đồng H_0) thì hai đại lượng đem so sánh được coi là khác nhau. Trong trường hợp so sánh được thực hiện bằng cách thiết lập tỷ số giữa các đại lượng cần so sánh, nếu như tỷ số của hai đại lượng cần so sánh nằm ở *miền lân cận giá trị trung bình (bằng 1) của phân phối (miền ổn định chấp nhận giả thuyết tương đồng H_0)* thì hai đại lượng đem so sánh có thể coi là như nhau. Trái lại, nếu tỷ số của hai đại lượng cần so sánh nằm ngoài miền này (miền ổn định bác bỏ giả thuyết tương đồng H_0) thì hai đại lượng đem so sánh được coi là khác nhau. Trong thống kê học mức *xác suất sai lầm $P = 0,05$* là *giới hạn sai lầm cho phép được dùng làm ranh giới phân chia hai miền: miền “nội tộc” và miền “ngoại tộc”*. Miền “nội tộc” là miền lân cận trị trung bình của phân phối, có thể đồng nhất với trị trung bình của phân phối, cho nên đây là miền ổn định chấp nhận giả thuyết tương đồng H_0 . Trái lại, miền “ngoại tộc” nằm cách xa trị trung bình của phân phối, không thể đồng nhất với trị trung bình của phân phối, cho nên đây là miền ổn định bác bỏ giả thuyết tương đồng H_0 . Như vậy, giá trị nào của phân phối nằm ở miền “ngoại tộc” (bác bỏ giả thuyết H_0) thì sẽ có xác suất sai lầm $P \leq 0,05$. Trái lại, giá trị nào của phân phối nằm ở miền “nội tộc” (chấp nhận giả thuyết H_0) thì sẽ có xác suất sai lầm $P > 0,05$, nghĩa là trong trường hợp này xác suất sai lầm P đã vượt quá giới hạn mức xác suất sai lầm cho phép. *Xác suất sai lầm P càng nhỏ thì khả năng bác bỏ giả thuyết tương đồng H_0 càng lớn và ngược lại, xác suất sai lầm P càng lớn thì khả năng chấp nhận giả thuyết tương đồng H_0 càng cao (hình 2.3).*



Hình 2.3. Đồ thị minh họa miền “nội tộc” và miền “ngoại tộc” ở phân phối hình chuông dạng chuẩn $N(0, 1)$

Qua sự trình bày trên đây, chúng ta có thể tóm tắt nội dung tiến trình thực hiện trắc nghiệm U như sau:

Bước 1: Đặt giả thuyết tương đồng H_0 .

Bước 2: Thiết lập biểu thức chứa các đại lượng cần so sánh dưới dạng hiệu số (ở đây là U).

Bước 3: Xác định phân phối của U (ở đây là phân phối bình thường dạng chuẩn).

Bước 4: Kiểm tra giả thuyết H_0 như sau:

Nếu $|U| \geq U_{0,05} = 1,96 \Rightarrow P \leq 0,05 \Rightarrow U$ thuộc miền “ngoại tộc” ($U \neq 0$), do đó giả thuyết H_0 bị bác bỏ.

Nếu $|U| < U_{0,05} = 1,96 \Rightarrow P > 0,05 \Rightarrow U$ thuộc “nội tộc” ($U \sim 0$), do đó giả thuyết H_0 được chấp nhận.

Bây giờ để minh họa chúng ta thử xem một vấn đề cụ thể sau đây: giả sử muốn so sánh một tỷ lệ $h = \frac{x}{n}$ với một tỷ lệ chuẩn đã có sẵn từ trước là p_0 , chúng ta cần tiến hành các bước như sau:

Bước 1: Đặt giả thuyết tương đồng $H_0: p = p_0$

Bước 2+3: Vì rằng $h = \frac{x}{n}$ là ước lượng của p của phân phối nhị thức dạng $B(n, p)$ nên h cũng có phân phối nhị thức với trung bình biến lượng như sau:

$$M[h] = M\left[\frac{x}{n}\right] = p$$

$$V[h] = V\left[\frac{x}{n}\right] = p \frac{(1-p)}{n}$$

Do đó, nếu $np \geq 5$ và $n(1-p) \geq 5$ thì h sẽ có phân phối cận bình thường dạng $N(p, \sqrt{\frac{p(1-p)}{n}})$ và $U = \frac{h-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{h-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ sẽ có phân phối cận bình thường dạng $N(0, 1)$.

Bước 4: Kiểm tra giả thuyết H_0 :

$$\text{Nếu } |U| \geq U_{0,05} = 1,96 \Rightarrow P \leq 0,05 \Rightarrow U \neq 0 \Rightarrow \frac{h-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \neq 0 \Rightarrow$$

$$\Rightarrow h - p_0 \neq 0 \Rightarrow h \neq p_0$$

$$\text{Nếu } |U| < U_{0,05} = 1,96 \Rightarrow P > 0,05 \Rightarrow U \sim 0 \Rightarrow \frac{h - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim 0 \Rightarrow$$

$$\Rightarrow h - p_0 \sim 0 \Rightarrow h \sim p_0$$

Ví dụ 2.1: Tỷ lệ phế phẩm theo quy định chung là 15%. Tiến hành kiểm tra 100 sản phẩm của nhà máy thấy có 10 phế phẩm. Kết luận như thế nào về điều này?

$$\text{Vì rằng: } |U| = \frac{\left| \frac{10}{100} - \frac{15}{100} \right|}{\sqrt{\frac{\frac{15}{100} \left(1 - \frac{15}{100} \right)}{100}}} \approx 1,4; P > 0,05.$$

Do đó, tỷ lệ phế phẩm của nhà máy trong đợt kiểm tra này là như quy định. Trong trường hợp này nếu như chúng ta không dùng trắc nghiệm U để đánh giá thì có lẽ chúng ta đã kết luận khác rồi (10% so với 15%). Thực ra không phải như vậy, kết quả trắc nghiệm U cho thấy rằng 10% chẳng qua chỉ là một “thăng giáng ngẫu nhiên” của 15%. Chúng ta sẽ xem xét vấn đề này chi tiết hơn ở chương 3 “so sánh các nghiệm thức”.

Sau đây là các trị thông dụng U_p của phân phối bình thường dạng chuẩn $N(0, 1)$ dùng cho trắc nghiệm U:

P(2 đuôi)	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
U_p	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

2.1.2.2. Khoảng tin tưởng và thuật so sánh dựa vào khoảng tin tưởng

Như chúng ta đã biết, nếu đại lượng ngẫu nhiên X có phân phối bình thường dạng $N(m, \sigma)$ thì $\bar{x}$ có phân phối bình thường dạng $N(m, \sqrt{\frac{\sigma^2}{n}})$ và

$$u = \frac{\bar{x} - m}{\sqrt{\frac{\sigma^2}{n}}} \text{ có phân phối bình thường dạng chuẩn } N(0, 1), \text{ do đó, khoảng tin}$$

tưởng (confidence interval) của trung bình của quần thể (m) ở mức xác suất

$$\text{sai lầm } P \text{ sẽ như sau: } -U_p < U < U_p \Leftrightarrow -U_p < \frac{\bar{x} - m}{\sqrt{\frac{\sigma^2}{n}}} < U_p \Leftrightarrow -U_p \sqrt{\frac{\sigma^2}{n}} < \bar{x}$$

- $m < U_p \sqrt{\frac{\sigma^2}{n}} \Leftrightarrow \bar{x} - U_p \sqrt{\frac{\sigma^2}{n}} < m < \bar{x} + U_p \sqrt{\frac{\sigma^2}{n}}$ hay $m \in (\bar{x} - U_p \sqrt{\frac{\sigma^2}{n}}; \bar{x} + U_p \sqrt{\frac{\sigma^2}{n}})$. Trong trường hợp phân phối đối xứng, khoảng tin

tương của m thường được ghi là: $m = \bar{x} \pm U_p \sqrt{\frac{\sigma^2}{n}}$ và $U_p \sqrt{\frac{\sigma^2}{n}}$ được gọi là *sai số cho phép* ở mức xác suất sai lầm P , thường được ký hiệu là δ . Chú ý rằng nếu $P = 0,05$ thì được coi là mặc nhiên không cần phải nêu ra, còn nếu như P ở các mức xác suất khác thì phải nêu rõ ra để tránh hiểu lầm.

Chú ý rằng phân phối bình thường là giới hạn của phân phối nhị thức cũng như phân phối Poisson khi $n \rightarrow \infty$, do đó, khi n lớn thì:

(i) Phân phối nhị thức có thể được coi như là một phân phối cận bình thường, do đó:

- $h = \frac{x}{n}$ có phân phối cận bình thường với trung bình và biến lượng như sau:

$$M[h] = M\left[\frac{x}{n}\right] = p$$

$$V[h] = V\left[\frac{x}{n}\right] = \frac{p(1-p)}{n} \approx \frac{h(1-h)}{n}$$

- $U = \frac{h-p}{\sqrt{\frac{h(1-h)}{n}}}$ có phân phối cận bình thường dạng $N(0, 1)$.

Do đó, khoảng tin tương của p ở mức xác suất sai lầm P sẽ như sau:

$$h - U_p \sqrt{\frac{h(1-h)}{n}} < p < h + U_p \sqrt{\frac{h(1-h)}{n}}$$

(ii) Phân phối Poisson có thể được coi như là một phân phối cận bình thường, do đó:

- X có phân phối cận bình thường dạng $N(\lambda, \sqrt{\lambda})$.

- $\bar{x}$ sẽ có phân phối cận bình thường với trung bình và biến lượng như sau:

$$M[\bar{x}] = \lambda$$

$$V[\bar{x}] = \frac{\lambda}{n} \approx \frac{\bar{x}}{n}$$

$$U = \frac{\bar{x} - \lambda}{\sqrt{\frac{\bar{x}}{n}}} \text{ có phân phối cận bình thường dạng } N(0, 1).$$

Do đó, khoảng tin tưởng của λ ở mức xác suất sai lầm P như sau:

$$\bar{x} - U_P \sqrt{\frac{\bar{x}}{n}} < \lambda < \bar{x} + U_P \sqrt{\frac{\bar{x}}{n}}$$

Khoảng tin tưởng có thể được vận dụng để thực hiện việc so sánh các nghiệm thức. Thuật so sánh dựa căn bản vào khoảng tin tưởng được tiến hành như sau:

Bước 1: Thiết lập biểu thức chứa các đại lượng cần so sánh dưới dạng hiệu số hoặc tỷ số.

Bước 2: Xác định luật phân phối.

Bước 3: Xác định khoảng tin tưởng. Nếu khoảng tin tưởng bao gồm trị zero (trường hợp các đại lượng cần so sánh nằm dưới dạng hiệu số) hoặc 1 (trường hợp các đại lượng cần so sánh nằm dưới dạng tỷ số) thì các đại lượng đem so sánh được coi là như nhau, ngược lại, nếu khoảng tin tưởng không bao gồm trị zero (hoặc 1) thì các đại lượng đem so sánh được coi là khác nhau.

Ví dụ 2.2. Tiến hành kiểm tra 100 sản phẩm của nhà máy thấy có 10 phế phẩm. Hãy xác định khoảng tin tưởng của tỷ lệ phế phẩm ở mức xác suất sai lầm $P = 0,05$.

$$\text{Vì rằng: } U = \frac{h - p}{\sqrt{\frac{h(1-h)}{n}}} = \frac{\frac{10}{100} - p}{\sqrt{\frac{\frac{10}{100} \left(1 - \frac{10}{100}\right)}{100}}} \text{ có phân phối cận bình thường}$$

dạng $N(0, 1)$, do đó, khoảng tin tưởng của p ở mức xác suất sai lầm $P = 0,05$, nghĩa là $U_{0,05} = 1,96$, sẽ như sau:

$$- 1,96 < \frac{0,1 - p}{\sqrt{\frac{0,1(1-0,1)}{100}}} < 1,96 \Leftrightarrow 0,1 - 1,96 \sqrt{\frac{0,1(1-0,1)}{100}} < p < 0,1 +$$

$$1,96 \sqrt{\frac{0,1(1-0,1)}{100}} \Leftrightarrow 0,1 - 0,0588 < p < 0,1 + 0,0588 \Leftrightarrow 0,0412 < p <$$

0,1588 hay $p = 0,1 \pm 0,0588$. Nhận xét: Vì rằng $P = 0,15 \in (0,0412; 0,1588)$, do

đó giả thuyết $H_0: P = 0,15$ được chấp nhận. Kết quả này hoàn toàn phù hợp với kết luận rút ra từ trắc nghiệm với giả thuyết tương đồng $H_0: P = 0,15$ đã thực hiện ở ví dụ 2.1.

Ví dụ 2.3. Tiến hành kiểm tra 100 sản phẩm của nhà máy A thấy có 8 phế phẩm. Ở nhà máy B, kiểm tra 300 sản phẩm thấy có 13 phế phẩm. Kết luận như thế nào về điều này.

Vì rằng h_1 có phân phối cận bình thường dạng $N(p_1, \sqrt{\frac{p(1-p)}{n_1}})$, h_2 có phân phối cận bình thường dạng $N(p_2, \sqrt{\frac{p(1-p)}{n_2}})$, $h_1 - h_2$ có phân phối cận bình thường dạng $N(p_1 - p_2, \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)})$ và $U = \frac{(h_1 - h_2) - (p_1 - p_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ có

phân phối cận bình thường dạng $N(0, 1)$ do đó, khoảng tin tưởng $P_1 - P_2$ ở mức xác suất sai lầm $P = 0,05$, nghĩa là $U_{0,05} = 1,96$, sẽ như sau:

$$\begin{aligned} -1,96 &< \frac{(h_1 - h_2) - (p_1 - p_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < 1,96 \Leftrightarrow (h_1 - h_2) \\ &- 1,96 \cdot \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} < p_1 - p_2 < (h_1 - h_2) + \\ &+ 1,96 \cdot \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ ở đây: } h_1 - h_2 = \frac{8}{10} - \frac{13}{300} \approx 0,0367 \\ p \approx h &= \frac{8+13}{100+300} = \frac{21}{400} \Rightarrow 1,96 \cdot \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \approx \\ &1,96 \cdot \sqrt{\frac{21}{400} \cdot \frac{379}{400} \cdot \left(\frac{1}{100} + \frac{1}{300}\right)} \approx 0,0505 \end{aligned}$$

do đó: $0,0367 - 0,0505 < p_1 - p_2 < 0,0367 + 0,0505$

Hay $-0,0138 < p_1 - p_2 < 0,0872$

Vì khoảng tin tưởng của $P_1 - P_2$ là $(-0,0138; 0,0872)$ bao hàm trị zero, nghĩa là $P_1 - P_2 \sim 0 \Rightarrow P_1 \sim P_2$. Điều này có nghĩa là tỷ lệ phế phẩm ở hai nhà máy A và B thực ra là như nhau. Kết quả này hoàn toàn phù hợp với kết luận rút ra từ trắc nghiệm U với giả thuyết tương đồng $H_0 : P_1 = P_2$ mà chúng ta sẽ thấy ở chương 3.

2.1.2.3. Ước định kích thước của mẫu

Như đã trình bày ở trên, khoảng tin tưởng của trung bình ở mức xác suất sai lầm P có thể được ghi dưới dạng như sau:

$$m = \bar{x} \pm U_p \cdot \sqrt{\frac{\sigma^2}{n}}$$

$$p = h \pm U_p \cdot \sqrt{\frac{h(1-h)}{n}}$$

$$\lambda = \bar{x} \pm U_p \cdot \sqrt{\frac{\bar{x}}{n}}$$

Với $U_p \cdot \sqrt{\frac{\sigma^2}{n}}$, $U_p \cdot \sqrt{\frac{h(1-h)}{n}}$, $U_p \cdot \sqrt{\frac{\bar{x}}{n}}$ được gọi là sai số cho phép ở mức

xác suất sai lầm P và thường được ký hiệu là δ . Vấn đề đặt ra ở đây là, với mong muốn đạt được một sai số cho phép không quá sai số ε quy định nào đó thì số số liệu (kích thước của mẫu, n) cần thiết phải lấy sẽ là bao nhiêu. Để xác định n chúng ta cần tiến hành giải các bất đẳng thức sau đây:

$$U_p \cdot \sqrt{\frac{\sigma^2}{n}} \leq \varepsilon, \quad U_p \cdot \sqrt{\frac{h(1-h)}{n}} \leq \varepsilon \text{ hoặc } U_p \cdot \sqrt{\frac{\bar{x}}{n}} \leq \varepsilon$$

Ví dụ 2.4. Kiểm tra 100 sản phẩm thấy có 10 phế phẩm. Phải tiến hành kiểm tra bao nhiêu sản phẩm để sai số cho phép của tỷ lệ phế phẩm không vượt quá 1% ở mức xác suất sai lầm $P = 0,05$.

$$\begin{aligned} \text{Vì rằng: } U_{0,05} \cdot \sqrt{\frac{h(1-h)}{n}} &= 1,96 \cdot \sqrt{\frac{10 \left(1 - \frac{10}{100}\right)}{n}} \leq \\ &\leq \frac{1}{100} \Rightarrow n \geq \frac{100}{100} \cdot \frac{90}{100} [(1,96)(100)]^2 = 3457,44 \end{aligned}$$

Vậy cần phải kiểm tra tối thiểu là $n = 3458$ sản phẩm.

2.1.2.4. Kiểm tra tính ngẫu nhiên của dãy số liệu

Để kiểm tra tính ngẫu nhiên của dãy số liệu $x_1, x_2, \dots, x_n$ chúng ta tiến hành như sau:

Bước 1: Sắp xếp dãy số liệu theo thứ tự từ nhỏ đến lớn và xác định trung vị (median) M_e của chúng.

Bước 2: Lập hiệu số $x_i - M_e$ để xác định tổng số dấu $+(n_1)$, tổng số dấu $-(n_2)$ và tổng số chuỗi (N).

Bước 3: Vì rằng:
$$U = \frac{\left(N - \frac{2n_1n_2}{n_1 + n_2} \right) - \frac{1}{2}}{\sqrt{\frac{2n_1n_2[2n_1n_2 - (n_1 + n_2)]}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}}$$
 có phân phối cận bình

thường dạng $N(0,1)$, do đó trắc nghiệm U sẽ được dùng để đánh giá tính ngẫu nhiên của dãy số liệu ở mức xác suất sai lầm P. Nếu $P \leq 0,05$ thì các số liệu thu được là không ngẫu nhiên, còn nếu $P > 0,05$ thì các số liệu thu được là ngẫu nhiên, nghĩa là, nếu U nằm ở miền “nội tộc” thì các số liệu là ngẫu nhiên (số liệu thu được là các thăng giáng ngẫu nhiên), ngược lại, nếu U nằm ở miền “ngoại tộc” thì các số liệu là không ngẫu nhiên (số liệu thu được không phải là các thăng giáng ngẫu nhiên).

Ví dụ 2.5. Tiến hành đo chiều cao một số cây ở vườn ươm kết quả như sau (cm) : 31, 39, 40, 45, 27, 28, 35, 55, 21, 33, 42, 36. Chúng ta thử xem số liệu thu được này có ngẫu nhiên hay không.

Trước hết chúng ta sắp xếp các số liệu theo thứ tự từ nhỏ đến lớn như sau: 21, 27, 28, 31, 33, 35, 36, 39, 40, 42, 45, 55. Do đó, trung vị sẽ là $M_e = \frac{(35+36)}{2} = 35,5$.

X_i	31	39	40	45	27	28	35	55	21	33	42	36
$x_i - M_e$	-	+	+	+	-	-	-	+	-	-	+	+

Tổng số dấu + là $n_1 = 6$, tổng số dấu - là $n_2 = 6$ và tổng số chuỗi là $N = 6$, do đó:

$$U = \frac{\left(6 - \frac{2 \cdot 6 \cdot 6}{6 + 6} \right) - \frac{1}{2}}{\sqrt{\frac{2 \cdot 6 \cdot 6 \cdot [2 \cdot 6 \cdot 6 - (6 + 6)]}{(6 + 6)^2 \cdot (6 + 6 - 1)}}} = -0,3028; P > 0,7$$

Kết quả trắc nghiệm U cho thấy rằng các số liệu thu được là ngẫu nhiên.

Chú ý rằng, trong trường hợp hiệu số $x_i - M_c = 0$, vì không mang dấu + hay - nên không được tính vào n_1 , n_2 hay N.

Trong trường hợp dãy số liệu chỉ gồm 2 biểu tượng như 0 và 1, + và - hoặc A và B ... thì n_1 , n_2 và N sẽ được xác định trực tiếp từ dãy số liệu thu được.

Ví dụ 2.6. Xác định tính ngẫu nhiên của dãy số liệu sau đây: 11001 00010 10100 11011 01001 00001 01000 1.

Vì rằng: Tổng số 0 là $n_1 = 21$, tổng số 1 là $n_2 = 15$ và tổng số chuỗi là $N = 23$, do đó:

$$U = \frac{\left(23 - \frac{2.21.15}{21+15}\right) - \frac{1}{2}}{\sqrt{\frac{2.21.15[2.21.15 - (21+15)]}{(21+15)^2.(21+15-1)}}} = 1,741; P > 0,08$$

Kết quả trắc nghiệm U cho thấy rằng các số liệu thu được là ngẫu nhiên.

2.2. PHÂN PHỐI $\chi^2(K)$

2.2.1. Đặc điểm

Nếu như có K đại lượng ngẫu nhiên độc lập U_i ($i = 1, 2, \dots, K$) và đều có phân phối bình thường dạng $N(0,1)$ thì tổng các bình phương: $U_1^2 + U_2^2 + \dots + U_K^2$ sẽ phân phối χ^2 với K độ tự do, tức là:

$$\chi^2(K) = \sum_{i=1}^K U_i^2 = U_1^2 + U_2^2 + \dots + U_K^2$$

P(*)	0,500	0,400	0,300	0,250	0,200	0,100	0,050	0,025	0,020	0,010	0,005	0,001
$\chi_p^2(1)$	0,455	0,71	1,07	1,32	1,64	2,71	3,84	5,02	5,41	6,63	7,88	10,83
U_p	0,674	0,842	1,036	1,149	1,282	1,645	1,960	2,241	2,326	2,576	2,807	3,291
U_p^2	0,455	0,71	1,07	1,32	1,64	2,71	3,84	5,02	5,41	6,63	7,88	10,83

(*) Trị xác suất sai lầm P ở phân phối χ^2 được tính từ 1 đuôi (1 phía) còn ở phân phối bình thường dạng $N(0,1)$ được tính từ 2 đuôi (2 phía).

Như vậy, nếu đại lượng ngẫu nhiên U của phân phối bình thường dạng $N(0,1)$ thì U^2 sẽ có phân phối χ^2 với 1 độ tự do, tức là:

$$\chi^2(1) = U^2$$

Ở trên là một vài trị thông dụng U_p của phân phối bình thường dạng $N(0,1)$ và $\chi_p^2(K)$ của phân phối $\chi^2(K)$ để minh họa cho mối quan hệ $\chi^2(1) = U^2$.

Như vậy rõ ràng là trắc nghiệm χ^2 với 1 độ tự do tương đương với trắc nghiệm U^2 , còn trắc nghiệm U tương đương với trắc nghiệm $\sqrt{\chi^2(1)}$.

Ngoài ra, trong trường hợp K lớn ($K > 30$) thì phân phối χ^2 có dạng gần như đối xứng, do đó, có thể được xác định thông qua phân phối bình thường dạng chuẩn $N(0,1)$ như sau:

$$- \chi_p^2(K) \approx \frac{1}{2} \left(U_{2p} + \sqrt{2K-1} \right)^2, \text{ nếu } P \leq 0,5$$

$$- \chi_p^2(K) \approx K \left(1 - \frac{2}{9K} - U_{2(1-p)} \cdot \sqrt{\frac{2}{9K}} \right)^3, \text{ nếu } P > 0,5$$

Chẳng hạn:

$$\chi_{0,05}^2(40) \approx \frac{1}{2} \left(1,645 + \sqrt{2 \cdot 40 - 1} \right)^2 \approx 55,47 \text{ (trị đúng là 55,8)}$$

$$\chi_{0,99}^2(40) \approx 40 \left(1 - \frac{2}{9 \cdot 40} - 2,326 \cdot \sqrt{\frac{2}{9 \cdot 40}} \right)^3 \approx 22,14 \text{ trị đúng là 22,2}.$$

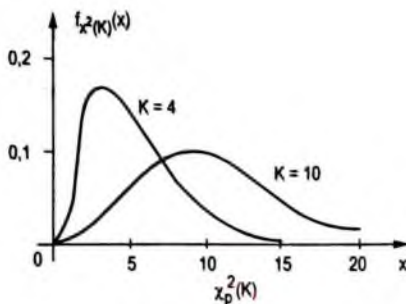
Phân phối $\chi^2(K)$ có hàm mật độ $f_{\chi^2(K)}(x)$ như sau:

$$f_{\chi^2(K)}(x) = \begin{cases} 0 & \text{nếu } x \leq 0 \\ \frac{1}{2^{\frac{K}{2}} G\left(\frac{K}{2}\right)} \cdot x^{\frac{K-2}{2}} \cdot e^{-\frac{x}{2}}, & \text{nếu } x > 0 \end{cases}$$

Ở đây $G\left(\frac{k}{2}\right)$ là hàm gamma của Euler*.

Đồ thị của hàm mật độ $f_{\chi^2(K)}(x)$ như sau (hình 2.4)

* Dạng tổng quát của hàm gamma của Euler như sau: $G(a) = \int_0^{+\infty} t^{a-1} \cdot e^{-t} dt$



Hình 2.4. Đồ thị hàm mật độ $f_{\chi^2(K)}(x)$ của phân phối $\chi^2(K)$

với $K = 4$ và $K = 10$

Phân phối $\chi^2(K)$ có trung bình là $M[\chi^2(K)] = K$ và biến lượng là $V[\chi^2(K)] = 2K$.

Chú ý rằng, trong thực tế thì phân phối $\chi^2(K)$ thường được xác định như sau: nếu như các số liệu $x_1, x_2, \dots, x_n$ của mẫu được lấy từ một quần thể có phân phối bình

thường dạng $N(m, \sigma)$ $\bar{x} = \sum_{i=1}^n \frac{x_i}{n}$ và $S^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(n-1)}$ là trung bình và biến lượng

của mẫu thì đại lượng $\frac{(n-1)S^2}{\sigma^2}$ sẽ có phân phối χ^2 với $n-1$ độ tự do, tức là:

$$\chi^2(n-1) = \frac{(n-1)S^2}{\sigma^2}$$

Từ đó chúng ta suy ra được các đẳng thức sau đây:

$$\frac{\chi^2(n-1)}{n-1} = \frac{S^2}{\sigma^2} \quad \text{hay} \quad \frac{\chi^2(K)}{K} = \frac{S^2}{\sigma^2} \quad \text{với } K = n-1 \text{ độ tự do.}$$

Điều này cũng dễ chứng minh như sau:

$$\chi^2(n-1) = \sum_{i=1}^{n-1} U_i^2 = \sum_{i=1}^{n-1} \left(\frac{x_i - m}{\sigma} \right)^2 = \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n-1} (x_i - m)^2 = \frac{1}{\sigma^2} \cdot (n-1)S^2$$

Chú ý rằng, đại lượng $\frac{S^2}{\sigma^2}$ có phân phối Fisher với K và ∞ độ tự do, tức là:

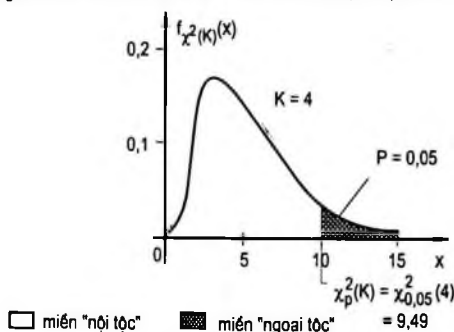
$$\frac{\chi^2(K)}{K} = F(k, \infty)$$

Sau đây là một vài trị thông dụng $\chi_p^2(K)$ của phân phối $\chi^2(K)$ và $F_p(K_1, K_2)$

của phân phối Fisher $F(K_1, K_2)$ để minh họa cho mối quan hệ $\frac{\chi^2(K)}{K} = F(K, \infty)$:

P(1 đuôi)	0,950	0,900	0,800	0,700	0,500	0,300	0,200	0,100	0,050	0,025	0,020	0,010	0,005	0,001
$F_p(5, \infty)$	0,23	0,32	0,49	0,60	0,87	1,21	1,46	1,85	2,21	2,57	2,68	3,02	3,35	4,10
$\chi_p^2(5)$	1,15	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,1	12,8	13,39	15,1	16,7	20,5
$\frac{\chi_p^2(5)}{5}$	0,23	0,32	0,49	0,60	0,87	1,21	1,46	1,85	2,21	2,57	2,68	3,02	3,35	4,10

Vì rằng phân phối $\chi^2(K)$ có trung bình là $M[\chi^2(K)] = K$, do đó, nếu như có một giá trị $\chi^2(K)$ nào đó nằm lân cận trị trung bình của phân phối, tức nằm ở miền “nội tộc” hay *miền ổn định chấp nhận giả thuyết tương đồng H_0* , thì có thể coi như $\chi^2(K) \sim K$ hay $\frac{\chi^2(K)}{K} \sim 1$, trái lại, nếu như giá trị $\chi^2(K)$ nằm xa trị trung bình của phân phối, tức nằm ở miền “ngoại tộc” hay *miền ổn định bác bỏ giả thuyết tương đồng H_0* , thì coi như $\chi^2(K) \neq K$ hay $\frac{\chi^2(K)}{K} \neq 1$. Như đã được trình bày ở trên, mức xác suất sai lầm $P = 0,05$ thường được dùng làm ranh giới phân chia giữa miền “nội tộc” và miền “ngoại tộc”. Các giá trị $\chi_p(K)$ nào nằm ở miền “ngoại tộc” (miền ổn định bác bỏ giả thuyết H_0) sẽ có mức xác suất sai lầm $P \leq 0,05$, trái lại, các giá trị $\chi_p^2(K)$ nào nằm ở miền “nội tộc” (miền ổn định chấp nhận giả thuyết H_0) sẽ có mức xác suất sai lầm $P > 0,05$ (hình 2.5).



Hình 2.5. Đồ thị minh họa miền “nội tộc” và miền “ngoại tộc” của phân phối $\chi^2(K)$ với $K = 4$

Trắc nghiệm so sánh dựa trên căn bản của phân phối $\chi^2(K)$ được gọi là “trắc nghiệm χ^2 ”. Để hiểu rõ bản chất cũng như mối quan hệ giữa “trắc nghiệm χ^2 ” và “trắc nghiệm F” (sẽ được trình bày chi tiết ở mục “phân phối Fisher”) chúng

ta hãy chú ý đến đẳng thức đã được trình bày ở trên là $\frac{\chi^2(K)}{K} = \frac{S^2}{\sigma^2}$ với K là độ

tự do và cũng là giá trị trung bình của phân phối $\chi^2(K)$, còn S^2 là biến lượng của mẫu và σ^2 là biến lượng của quần thể có phân phối bình thường dạng $N(m, \sigma)$. Điều này có nghĩa là, nếu S^2 (với $n - 1 = K$ độ tự do) là ước lượng tốt nhất của σ^2

của phân phối bình thường dạng $N(m, \sigma)$ thì đại lượng $\frac{KS^2}{\sigma^2}$ sẽ có phân phối χ^2

với K độ tự do, $\chi^2(K)$, và đại lượng $\frac{S^2}{\sigma^2}$ sẽ có phân phối Fisher với K và ∞ độ tự

do, $F(K, \infty)$. Để nhận thấy là, nếu n (hay K) $\rightarrow \infty$ thì $S^2 \rightarrow \sigma^2$ khi đó

$\frac{\chi^2(K)}{K} = \frac{S^2}{\sigma^2} \rightarrow 1$, tức là $\chi^2(K) \rightarrow K$ (thuộc miền “nội tộc”). Như vậy, nhờ mối

quan hệ với tỷ số biến lượng $\frac{S^2}{\sigma^2}$ mà chúng ta hiểu được bản chất của trắc

nh nghiệm χ^2 như sau: Giả sử biểu thức chứa các nội dung cần trắc nghiệm là χ^2 (K) có phân phối χ^2 thì:

- Nếu $\chi^2(K)$ nằm ở miền “nội tộc” của phân phối (miền ấn định chấp nhận giả thuyết tương đồng H_0), nghĩa là $\chi^2(K)$ nằm lân cận giá trị trung bình của phân phối, tức $\chi^2(K) \sim K$, thì giả thuyết H_0 được chấp nhận:

$$\chi^2(K) \sim K \Leftrightarrow \frac{\chi^2(K)}{K} \sim 1 \Leftrightarrow \frac{S^2}{\sigma^2} \sim 1 \Leftrightarrow S^2 \sim \sigma^2$$

- Nếu $\chi^2(K)$ nằm ở miền “ngoại tộc” của phân phối (miền ấn định bác bỏ giả thuyết tương đồng H_0), nghĩa là $\chi^2(K)$ nằm xa giá trị trung bình của phân phối, tức $\chi^2(K) \neq K$, thì giả thuyết H_0 bị bác bỏ:

$$\chi^2(K) \neq K \Leftrightarrow \frac{\chi^2(K)}{K} \neq 1 \Leftrightarrow \frac{S^2}{\sigma^2} \neq 1 \Leftrightarrow S^2 \neq \sigma^2$$

Về phương diện thực hành giá trị χ^2 thường được xác định bởi công thức sau đây:

$$\chi^2 = \sum \frac{(\text{Trị thực tế} - \text{Trị lý thuyết})^2}{\text{Trị lý thuyết}}$$

Bây giờ chúng ta tiến hành chứng minh biểu thức trên để có thể hiểu rõ hơn bản chất của trắc nghiệm χ^2 .

Giả sử rằng xác suất xảy ra sự kiện Δ_i của biến ngẫu nhiên x_i là p_i với $\sum_{i=1}^r p_i = 1$. Như vậy rõ ràng là các xác suất p_i sẽ rất nhỏ, do đó, x_i sẽ có phân phối Poisson với trung bình và biến lượng là $M[x_i] = V[x_i] = \lambda_i$.

	Δ_1	Δ_2	...	Δ_r	Tổng
Trị thực tế	x_1	x_2	...	x_r	n
Trị lý thuyết	np_1	np_2	...	np_r	n

Trong trường hợp n lớn hay $np_i \geq 5$ thì x_i sẽ có phân phối cận bình thường dạng $N(np_i, \sqrt{np_i})$ và $U_i = \frac{x_i - np_i}{\sqrt{np_i}}$ sẽ có phân phối cận bình thường dạng $N(0, 1)$, do đó:

$$\chi^2 = \sum_{i=1}^r U_i^2 = \sum_{i=1}^r \frac{(x_i - np_i)^2}{np_i}$$

Chú ý rằng, độ tự do của phân phối χ^2 trong trường hợp này phụ thuộc rất nhiều vào các giá trị xác suất P_i được dùng để tính các trị số lý thuyết np_i , nên không bằng r mà sẽ nhỏ hơn r . Về phương diện lý thuyết người ta đã chứng minh được rằng độ tự do của phân phối χ^2 trong trường hợp này là $K = r - l - 1$ với l là tham số chưa biết của phân phối lý thuyết được dùng để tính các giá trị xác suất P_i . Như vậy, nếu như các giá trị xác suất p_i được tính từ các phân phối lý thuyết không chứa tham số chưa biết, chẳng hạn như phân phối 3 : 1 và 9 : 3 : 3 : 1 của định luật di truyền Mendel, phân phối 1 : 1 về giới tính ở người..., thì $l = 0$. Trong trường hợp các giá trị xác suất p_i được tính từ phân phối nhị thức với một tham số p chưa biết, nghĩa là các trị xác suất P_i được tính nhờ ước lượng $\bar{p}$ của p , thì $l = 1$, ngược lại, nếu như các trị xác suất p_i được tính từ phân phối nhị thức với tham số p đã biết thì $l = 0$. Tương tự như vậy, đối với phân phối Poisson thì $l = 1$ nếu như tham số λ của phân phối là chưa biết, nghĩa là các trị xác suất P_i được tính nhờ ước lượng $\bar{\lambda}$ của λ , ngược lại thì $l = 0$ nếu như các trị xác suất P_i được tính từ phân phối Poisson với tham số λ đã biết. Trong trường hợp các trị xác suất P_i được tính

từ phân phối bình thường với cả hai tham số đều chưa biết là m và σ , nghĩa là các trị xác suất P_i được tính nhờ ước lượng $\bar{x}$ của m và ước lượng S của σ , thì $l = 2$, ngược lại thì l có thể bằng 1 (nếu chỉ có m hoặc σ là chưa biết) hay bằng zero (nếu cả hai tham số m và σ đều đã biết). Trong trường hợp của bảng $m \times n$, vì rằng phân phối lý thuyết không chứa tham số chưa biết nên các trị l đều bằng zero, do đó, độ tự do của phân phối χ^2 trong trường hợp này là $K = (m - 1)(n - 1)$.

2.2.2. Ứng dụng

2.2.2.1. Kiểm tra tính độc lập của hai đại lượng ngẫu nhiên

Hai đại lượng ngẫu nhiên X và Y được coi là *độc lập* nhau nếu như đại lượng này thay đổi không kéo theo sự thay đổi của đại lượng kia và ngược lại, nghĩa là mối quan hệ "nhân - quả" giữa X và Y không tồn tại, thường được ký hiệu là $X \nleftrightarrow Y$. Trái lại, hai đại lượng ngẫu nhiên X và Y được coi là *không độc lập* nhau nếu như đại lượng này thay đổi thì sẽ kéo theo sự thay đổi của đại lượng kia và ngược lại, nghĩa là mối quan hệ "nhân - quả" giữa X và Y tồn tại, thường được ký hiệu là $X \leftrightarrow Y$. Trắc nghiệm χ^2 thường được dùng để kiểm tra tính độc lập của hai đại lượng ngẫu nhiên X và Y như sau:

$\begin{matrix} Y \\ X \end{matrix}$	y_1	y_2	...	y_v	$\sum_{i=1}^v n_{ij} = n_i$
x_1	n_{11}	n_{12}	...	n_{1v}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2v}	$n_{2.}$
...	...	...	...	...	...
x_u	n_{u1}	n_{u2}	...	n_{uv}	$n_{u.}$
$\sum_{i=1}^v n_{ij} = n_j$	$n_{.1}$	$n_{.2}$	...	$n_{.v}$	$n_{..} = n$

Với giả thuyết $H_0 : X \nleftrightarrow Y$, ta có:

$$P_j = P[X = x_i; Y = y_j] = P[X = x_i][Y = y_j] = p_i p_j$$

Chú ý rằng, ước lượng của p_i là $\bar{p}_i = \frac{n_{i.}}{n}$ và ước lượng của p_j là $\bar{p}_j = \frac{n_{.j}}{n}$,

do đó, trị lý thuyết $\tilde{n}_{ij}$ sẽ là: $\tilde{n}_{ij} = np_{ij} = np_i p_j = n \left(\frac{n_{i.}}{n} \right) \left(\frac{n_{.j}}{n} \right) = \frac{n_{i.} n_{.j}}{n}$.

Do đó, giá trị χ^2 sẽ là:
$$\chi^2 = \sum_{i=1}^u \sum_{j=1}^v \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}$$

Với $K = (u - 0 - 1)(v - 0 - 1) = (u - 1)(v - 1)$ độ tự do.

Từ giá trị χ^2 tính được này ta xác định được xác suất sai lầm P và kiểm tra giả thuyết $H_0: X \leftrightarrow Y$ như sau:

- Nếu $P > 0,05$, nghĩa là trị $\chi_p^2(K)$ nằm ở miền “nội tộc” (miền ấn định chấp nhận giả thuyết tương đồng H_0) thì X và Y độc lập nhau ($X \leftrightarrow Y$).

- Nếu $P \leq 0,05$, nghĩa là trị $\chi_p^2(K)$ nằm ở miền “ngoại tộc” (miền ấn định bác bỏ giả thuyết tương đồng H_0) thì X và Y không độc lập nhau ($X \leftrightarrow Y$).

Chú ý rằng, cũng như trường hợp của trắc nghiệm U, khi dùng trắc nghiệm χ^2 cần đảm bảo điều kiện là trị lý thuyết $\tilde{n}_{ij} = np_{ij} = \frac{n_i n_j}{n} \geq 5$, trong trường hợp ngược lại thì cần gộp chung một số hàng cũng như một số cột lại với nhau cho đạt yêu cầu. Riêng đối với trường hợp của bảng 2×2 , vì không thể gộp chung các hàng cũng như các cột được nữa, do đó, hoặc là dùng “tu chỉnh Yates” (giảm bớt 0,5 cho mỗi giá trị tuyệt đối của hiệu số giữa trị thực tế và trị lý thuyết) hoặc là thay n bằng $n - 1$ trong công thức tính nhanh giá trị χ^2 . Như vậy, giá trị χ^2 tu chỉnh (χ_c^2) luôn nhỏ hơn giá trị χ^2 ban đầu, $\chi_c^2 < \chi^2$, do đó, xác suất sai lầm được tính từ trị χ_c^2 luôn luôn lớn hơn xác suất sai lầm được tính từ χ^2 ban đầu, $P[\chi_c^2] > P[\chi^2]$. Do đó, nếu như $P[\chi^2] > 0,05$ thì không cần phải tiến hành tu chỉnh cho χ^2 nữa.

Ví dụ 2.7. Thực hiện điều trị bệnh theo hai phương pháp khác nhau (A và B). Kết quả như sau:

Kết quả	Phương pháp điều trị		Tổng
	A	B	
Khỏi bệnh	14	24	38
Không khỏi bệnh	26	36	62
Tổng	40	60	100

Hãy cho biết kết quả điều trị của hai phương pháp này có khác nhau hay không.

Trước hết cần thảo luận về cách đặt giả thuyết tương đồng H_0 . Giả sử P_A và P_B là tỷ lệ khỏi bệnh từ phương pháp điều trị A và B thì giả thuyết tương

lồng có thể đặt là $H_0 : P_A = P_B$. Bây giờ chúng ta có thể xem biến ngẫu nhiên X là kết quả điều trị, còn biến ngẫu nhiên Y là phương pháp điều trị thì giả thuyết tương đồng có thể đặt là $H_0 : X \leftrightarrow Y$. Hai cách đặt giả thuyết tương đồng này về mặt hình thức thì trông có vẻ khác nhau nhưng nội dung (bản chất) thì hoàn toàn giống nhau. Thực vậy, vì rằng nếu biến ngẫu nhiên X thay đổi (khỏi bệnh $\neq$ không khỏi bệnh) mà không kéo theo sự thay đổi của biến ngẫu nhiên $Y (P_A = P_B)$ thì rõ ràng là X và Y là hai biến ngẫu nhiên độc lập nhau ($X \leftrightarrow Y$).

Bây giờ chúng ta xem xét đến kỹ thuật tính toán. Thông thường, giá trị χ^2 được tính từ công thức định nghĩa như sau:

Trị thực tế (n_{ij})	Trị lý thuyết ($\bar{n}_{ij}$)
$n_{11} = 14$	$\bar{n}_{11} = 40.38/100 = 15,2$
$n_{12} = 24$	$\bar{n}_{12} = 60.38/100 = 22,8$
$n_{21} = 26$	$\bar{n}_{21} = 40.62/100 = 24,8$
$n_{22} = 36$	$\bar{n}_{22} = 60.62/100 = 37,3$
$n = 100$	$n = 100$

$$\chi^2 = \frac{(14-15,2)^2}{15,2} + \frac{(24-22,8)^2}{22,8} + \frac{(26-24,8)^2}{24,8} + \frac{(36-37,2)^2}{37,2} = 0,255$$

Với $K = (2 - 1)(2 - 1) = 1$ độ tự do: $P > 0,5$. Vậy giả thuyết H_0 được chấp nhận, nghĩa là $X \leftrightarrow Y$ hay $P_A = P_B$, tức hiệu quả điều trị của hai phương pháp A và B là như nhau.

Thực ra, chúng ta có thể tiến hành tính nhanh trị χ^2 như sau:

$$\chi^2 = 100 \left(\frac{14^2}{40.38} + \frac{24^2}{60.38} + \frac{26^2}{40.62} + \frac{36^2}{60.62} - 1 \right) = 0,255$$

Ngoài ra, đối với bảng 2×2 chúng ta có thể tiến hành tính nhanh trị χ^2 như sau:

$$\chi^2 = \frac{100(14.36 - 26.24)^2}{40.60.38.62} = 0,255$$

Chú ý là, vì rằng $\chi^2(1) = U^2$, do đó, đối với các trắc nghiệm χ^2 với 1 độ tự do có thể được thay thế bởi trắc nghiệm U như sau (với giả thuyết $H_0 : P_A = P_B$).

$$U = \frac{(h_A - h_B) - (P_A - P_B)}{\sqrt{h(1-h) \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}} = \frac{\left(\frac{14}{40} - \frac{24}{60} \right) - 0}{\sqrt{\frac{38}{100} \left(1 - \frac{38}{100} \right) \left(\frac{1}{40} + \frac{1}{60} \right)}} = 0,5046; P > 0,5$$

Rõ ràng là $U^2 = 0,5046^2 = 0,255 = \chi^2(1)$

Ví dụ 2.8: Kết quả điều tra về bệnh mù màu như sau:

	Nam	Nữ	Tổng
Mù màu	38	6	44
Không mù màu	442	515	956
Tổng	480	520	1000

$$H_0 = X \leftrightarrow Y, \min[\bar{n}_{ij}] = 480 \cdot \frac{44}{1000} = 21,12 > 5$$

$$\chi^2 = \frac{1000(38 \cdot 515 - 442 \cdot 6)^2}{480 \cdot 520 \cdot 44 \cdot 956} = 27,14 \text{ với } K = (2 - 1)(2 - 1) = 1 \text{ độ tự do;}$$

$P < 0,001$. Vậy giả thuyết H_0 bị bác bỏ, nghĩa là X và Y không độc lập nhau ($X \leftrightarrow Y$), tức bệnh mù màu phụ thuộc vào giới tính.

Ví dụ 2.9: Kết quả điều tra về màu tóc ở các miền địa lý khác nhau (A, B và C) như sau:

Miền	Màu tóc			Tổng
	Vàng	Bạch kim	Hung hung	
A	2	9	9	20
B	3	6	21	30
C	15	15	20	50
Tổng	20	30	50	100

$$H_0 = X \leftrightarrow Y, \min[\bar{n}_{ij}] = 20 \cdot \frac{20}{1000} = 4 < 5$$

$$\chi^2 = 100 \left(\frac{2^2}{20 \cdot 20} + \frac{9^2}{30 \cdot 20} + \frac{9^2}{50 \cdot 20} + \frac{3^2}{20 \cdot 30} + \frac{6^2}{30 \cdot 30} + \frac{21^2}{50 \cdot 30} + \frac{15^2}{30 \cdot 50} + \frac{20^2}{50 \cdot 50} - 1 \right)$$

=

$$= 11 \text{ với } K = (3 - 1)(3 - 1) = 4 \text{ độ tự do; } P < 0,05.$$

Vì rằng $\min[\bar{n}_{ij}] = 4 < 5$ và $P < 0,05$, do đó cần phải tiến hành tu chỉnh giá trị χ^2 bằng cách gộp chung hàng A và B lại với nhau như sau:

Miền	Màu tóc			Tổng
	Vàng	Bạch kim	Hung hung	
AB	5	15	30	50
C	15	15	20	50
Tổng	20	30	50	100

$$H_0 = X \leftrightarrow Y, \min[\bar{n}_{ij}] = 20 \cdot \frac{50}{1000} = 10 > 5$$

$$\chi^2 = 100 \left(\frac{5^2}{20 \cdot 50} + \frac{15^2}{30 \cdot 50} + \frac{30^2}{50 \cdot 50} + \frac{15^2}{20 \cdot 50} + \frac{15^2}{30 \cdot 50} + \frac{20^2}{50 \cdot 50} - 1 \right) = 7$$

với $K = (2 - 1)(3 - 1) = 2$ độ tự do; $P < 0,05$. Vậy giả thiết H_0 bị bác bỏ, nghĩa là X và Y không độc lập nhau ($X \leftrightarrow Y$), tức màu tóc phụ thuộc vào các miền địa lý khác nhau.

2.2.2.2. Kiểm tra sự phù hợp của một phân phối

2.2.2.2.(a). Ứng với các phân phối thực nghiệm khoa học

Ví dụ 2.(10): Thực hiện việc lai giống cà chua người ta thu được 310 cà chua quả đỏ và 90 cà chua quả vàng. Hỏi kết quả này có nghiệm đúng tỷ lệ 3 : 1 theo định luật di truyền của Mendel không?

Trị thực tế (x_i)

$$x_1 = 310$$

$$x_2 = 90$$

$$n = 400$$

Trị lý thuyết (np_i)

$$np_1 = 400 \left(\frac{3}{4} \right) = 300$$

$$np_2 = 400 \left(\frac{1}{4} \right) = 100$$

$$n = 400$$

$$H_0: p_1 = \frac{3}{4}, p_2 = \frac{1}{4}; \min[np_i] = 100 > 5$$

$$\chi^2 = \frac{(310 - 300)^2}{300} + \frac{(90 - 100)^2}{100} = 1,33 \text{ với } K = r - 1 - 1 = 2 - 0 - 1 = 1 \text{ độ}$$

tự do; $P > 0,2$. Vậy giả thiết H_0 được chấp nhận, nghĩa là kết quả thu được nghiệm đúng tỷ lệ 3 : 1.

Thực ra trong trường hợp này chúng ta có thể tiến hành *tính nhanh giá trị* χ^2 như sau:

$$\chi^2 = \frac{\left[300 - 90\left(\frac{3}{1}\right)\right]^2}{400\left(\frac{3}{1}\right)} = 1,33$$

hay

$$\chi^2 = \frac{\left[90 - 300\left(\frac{1}{3}\right)\right]^2}{400\left(\frac{1}{3}\right)} = 1,33$$

Ngoài ra, như chúng ta đã biết là $\chi^2(1) = U^2$, do đó, có thể dùng trắc nghiệm U thay vì dùng trắc nghiệm χ^2 như sau (với giả thuyết $H_0 : p = 3/4$):

$$U = \frac{h - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{310}{400} - \frac{3}{4}}{\sqrt{\frac{\frac{3}{4}\left(1 - \frac{3}{4}\right)}{400}}} = 1,1547; P > 0,2$$

Rõ ràng là $U^2 = 1,1547^2 = 1,33 = \chi^2(1)$

Ví dụ 2.(11): Thực hiện việc lai giống bắp người ta thu được 773 bắp xanh, 231 bắp vàng, 238 bắp xanh sọc dưa và 59 bắp xanh sọc dưa vàng. Hỏi kết quả này có nghiệm đúng tỷ lệ 9 : 3 : 3 : 1 theo định luật di truyền của Mendel không ?

Trị thực tế (x_i)

$$x_1 = 733$$

$$x_2 = 231$$

$$x_3 = 238$$

$$x_4 = 59$$

$$n = 1301$$

Trị lý thuyết (np_i)

$$np_1 = 1301\left(\frac{9}{16}\right) = 731,9$$

$$np_2 = 1301\left(\frac{3}{16}\right) = 243,9$$

$$np_3 = 1301\left(\frac{3}{16}\right) = 243,9$$

$$np_4 = 1301\left(\frac{1}{16}\right) = 81,3$$

$$n = 1301$$

$$H_0: p_1 = \frac{9}{16}, p_2 = p_3 = \frac{3}{16}; p_4 = \frac{1}{16} \quad \min[np_i] = 81,3 > 5$$

$$\chi^2 = \frac{(773 - 731,9)^2}{731,9} + \frac{(231 - 243,9)^2}{243,9} + \frac{(238 - 243,9)^2}{243,9} + \frac{(59 - 81,3)^2}{81,3} = 9,25$$

với $K = r - l - 1 = 4 - 0 - 1 = 3$ độ tự do; $P < 0,05$. Vậy giả thuyết H_0 bị bác bỏ, nghĩa là kết quả thu được không nghiệm đúng tỷ lệ 9 : 3 : 3 : 1.

Ví dụ 2.(12): Để biết cấu trúc của đồng xu có đối xứng hay không người ta gieo đồng xu trong 50 lần thì thấy xuất hiện mặt sấp 20 lần. Kết luận thế nào về vấn đề này ?

Rõ ràng là, nếu đồng xu có cấu trúc đối xứng thì kết quả phải nghiệm đúng tỷ lệ 1 : 1.

Trị thực tế (x_i)

$$x_1 = 20$$

$$x_2 = 30$$

$$n = 50$$

Trị lý thuyết (np_i)

$$np_1 = 50 \cdot \frac{1}{2} = 25$$

$$np_2 = 50 \cdot \frac{1}{2} = 25$$

$$n = 50$$

$$H_0: p_1 = p_2 = \frac{1}{2}; \quad \min[np_i] = 25 > 5$$

$$\chi^2 = \frac{(20 - 25)^2}{25} + \frac{(30 - 25)^2}{25} = 2 \quad \text{với } K = r - l - 1 = 2 - 0 - 1 = 1 \text{ độ tự do;}$$

$P > 0,1$. Vậy giả thuyết H_0 được chấp nhận, nghĩa là kết quả gieo đồng xu nghiệm đúng tỷ lệ 1 : 1, do đó, đồng xu có cấu trúc đối xứng.

Thực ra trong trường hợp này chúng ta có thể tiến hành *tính nhanh giá trị χ^2* như sau:

$$\chi^2 = \frac{\left[20 - 30 \cdot \frac{1}{2} \right]^2}{50 \cdot \frac{1}{2}} = 2 \quad \text{hay} \quad \chi^2 = \frac{\left[30 - 20 \cdot \frac{1}{2} \right]^2}{50 \cdot \frac{1}{2}} = 2$$

Ngoài ra như chúng ta đã biết là các trắc nghiệm χ^2 với 1 độ tự do có thể được thay thế bởi trắc nghiệm U như sau (Với giả thuyết $H_0: p = \frac{1}{2}$):

$$U = \frac{h-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{20}{50} - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}\left(1-\frac{1}{2}\right)}{50}}} = -\sqrt{2}; P > 0,1$$

Rõ ràng là: $U^2 = (-\sqrt{2})^2 = 2 = \chi^2(1)$

2.2.2.2.(b). Ứng với các phân phối trong thống kê học

Ví dụ 2.(13): Thống kê số heo đực con trong mỗi lứa "lúa - 8 - con" cho kết quả sau đây:

Số heo đực con trong mỗi lứa "lúa - 8 - con"	0	1	2	3	4	5	6	7	8
Số "lúa - 8 - con" (x_i)	0	5	9	22	25	26	14	4	1

Hãy xem xét giới tính ở heo có tuân theo luật phân phối nhị thức $B(n, p = \frac{1}{2})$ hay không ?

Trị thực tế (x_i)

Trị lý thuyết (np_i)

$$\left. \begin{array}{l} X_0 = 0 \\ X_1 = 5 \\ x_2 = 9 \end{array} \right\} 14 \quad \left. \begin{array}{l} np_0 = 106.C_8^0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^8 = 0,414 \\ np_1 = 106.C_8^1 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^7 = 3,312 \\ np_2 = 106.C_8^2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^6 = 11,594 \end{array} \right\} 15,320$$

$$x_3 = 22 \quad np_3 = 106.C_8^3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^5 = 23,188$$

$$x_4 = 25 \quad np_4 = 106.C_8^4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^4 = 28,984$$

$$x_5 = 26 \quad np_5 = 106.C_8^5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^3 = 23,188$$

$$\begin{array}{l}
 \left. \begin{array}{l} x_6 = 14 \\ x_7 = 4 \\ x_8 = 1 \end{array} \right\} 19 \\
 \hline n = 106
 \end{array}
 \quad
 \begin{array}{l}
 np_6 = 106 \cdot C_8^6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^2 = 11,594 \\
 np_7 = 106 \cdot C_8^7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^1 = 3,312 \\
 np_8 = 106 \cdot C_8^8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^0 = 0,414 \\
 \hline n = 106
 \end{array}
 \quad
 \left. \begin{array}{l} \\ \\ \end{array} \right\} 15,320$$

$$H_0: p = \frac{1}{2}; \min[np_i] = 15,320 > 5$$

$$\begin{aligned}
 \chi^2 &= \frac{(14 - 15,320)^2}{15,320} + \frac{(22 - 23,188)^2}{23,188} + \frac{(25 - 28,984)^2}{28,984} + \frac{(26 - 23,188)^2}{23,188} + \\
 &+ \frac{(19 - 15,320)^2}{15,320} = 1,948 \text{ với } K = r - l - 1 = 5 - 0 - 1 = 4 \text{ độ tự do; } P > 0,5. \text{ Vậy} \\
 &\text{giả thuyết } H_0 \text{ được chấp nhận, nghĩa là giới tính ở heo tuân theo luật phân phối} \\
 &\text{nhị thức } B(n, p = \frac{1}{2}).
 \end{aligned}$$

Ví dụ 2.(14): Xử lý nảy mầm cho 800 hạt giống được thực hiện trong 80 khay, mỗi khay 10 hạt. Kết quả thu được như sau:

Số hạt nảy mầm ở mỗi khay	0	1	2	3	4	5	≥ 6
Số khay (x_i)	6	20	28	12	8	6	0

Hãy xét xem sự nảy mầm của hạt giống có tuân theo luật phân phối nhị thức $B(n, p)$ hay không ?

Chú ý rằng ở đây tham số p của phân phối là chưa biết, do đó $l = 1$ và ước lượng $\hat{p}$ của p được tính như sau:

$$\hat{p} = \frac{0.6 + 1.20 + 2.28 + 3.12 + 4.8 + 5.6 + 6.0}{800} = \frac{174}{800} = 0,2175$$

Trị thực tế (x_i)

Trị lý thuyết (np_i)

$$x_0 = 6$$

$$np_0 = 80 \cdot C_8^5 (0,2175)^0 (0,7825)^{10} = 6,885$$

$$x_1 = 20$$

$$np_1 = 80 \cdot C_{10}^1 (0,2175)^1 (0,7825)^9 = 19,139$$

$$x_2 = 28$$

$$np_2 = 80 \cdot C_{10}^2 (0,2175)^2 (0,7825)^8 = 23,938$$

$$x_3 = 12$$

$$np_3 = 80 \cdot C_{10}^3 (0,2175)^3 (0,7825)^7 = 17,744$$

$$\left. \begin{array}{l} x_4 = 8 \\ x_5 = 6 \\ \sum_{i=6}^{10} x_i = 0 \end{array} \right\} 14 \quad \left. \begin{array}{l} np_4 = 80 \cdot C_{10}^4 (0,2175)^4 (0,7825)^6 = 8,631 \\ np_5 = 80 \cdot C_{10}^5 (0,2175)^5 (0,7825)^5 = 2,879 \\ n \sum_{i=6}^{10} p_i = 80 - (6,885 + 19,139 + \dots + 2,879) = 0,784 \end{array} \right\} 12,294$$

$$n = 80 \qquad n = 80$$

$$H_0 : p = 0,2175; \min[np_i] = 6,885 > 5$$

$$\chi^2 = \frac{(6-6,885)^2}{6,885} + \frac{(20-19,139)^2}{19,139} + \frac{(28-23,938)^2}{23,938} + \frac{(12-17,744)^2}{17,744} + \frac{(14-12,294)^2}{12,294} = 2,938 \text{ với } K = r - l - 1 = 5 - 10 - 1 = 3 \text{ độ tự do; } P >$$

0,25. Vậy giả thuyết H_0 được chấp nhận, nghĩa là sự nảy mầm của hạt giống tuân theo luật phân phối nhị thức $B(n, p = 0,2175)$.

Ví dụ 2.(15): Thống kê số lần hỏng hóc của máy trong 10000 giờ hoạt động cho kết quả như sau:

Số lần hỏng hóc của máy	0	1	2	3	4	5	≥ 6
Số trường hợp xảy ra (x_i)	427	235	72	21	1	1	0

Hãy xét xem sự hỏng hóc của máy có tuân theo luật phân phối Poisson hay không ?

Chú ý rằng ở đây tham số λ của phân phối là chưa biết, do đó $l = 1$ và ước lượng $\tilde{\lambda}$ của λ được tính như sau:

$$\tilde{\lambda} = \frac{0.427 + 1.235 + 2.72 + 3.21 + 4.1 + 5.1 + 6.0}{427 + 235 + 72 + 21 + 1 + 1 + 0} = \frac{451}{757} = 0,6$$

(Dễ nhận thấy là $\tilde{\lambda} = n\bar{p} \Rightarrow 0,6 = 757\bar{p} \Rightarrow \bar{p} = \frac{6}{7570} \Rightarrow$ xác suất xảy ra hỏng hóc của máy là rất nhỏ).

Trị thực tế (x_i)

Trị lý thuyết (np_i)

$$x_0 = 427$$

$$np_0 = 757 \cdot \frac{0,6^0}{0!} e^{-0,6} = 416$$

$$x_1 = 235$$

$$np_1 = 757 \cdot \frac{0,6^1}{1!} e^{-0,6} = 249$$

$$x_2 = 72$$

$$np_2 = 757 \cdot \frac{0,6^2}{2!} e^{-0,6} = 75$$

$$x_3 = 21$$

$$np_3 = 757 \cdot \frac{0,6^3}{3!} e^{-0,6} = 15$$

$$x_4 = 1$$

$$np_4 = 757 \cdot \frac{0,6^4}{4!} e^{-0,6} = 2$$

$$x_5 = 1$$

$$np_5 = 757 \cdot \frac{0,6^5}{5!} e^{-0,6} = 0$$

$$\sum_{i=6}^{\infty} x_i = 0$$

$$\sum_{i=6}^{\infty} np_i = 0$$

$$n = 757$$

$$n = 757$$

$$H_0 : \lambda = 0,6; \min [np_i] = 17 > 5$$

$$\chi^2 = \frac{(427 - 416)^2}{416} + \frac{(235 - 249)^2}{249} + \frac{(72 - 75)^2}{75} + \frac{(23 - 17)^2}{17} = 3,316$$

với $K = r - l - 1 = 4 - 1 - 1 = 2$ độ tự do; $P > 0,1$. Vậy giả thuyết H_0 được chấp nhận, nghĩa là sự hỏng hóc của máy tuân theo luật phân phối Poisson với tham số $\lambda = 0,6$.

Ví dụ 2.(16): Đo chiều cao của 55 cây rừng cho kết quả sau đây:

20,3	13,9	19,9	13,5	18,3	20,4	17,8	22,8	18,1	16,7	15,3
15,3	19,5	17,8	17,2	18,5	14,7	11,8	14,5	19,7	17,5	12,5
14,3	17,8	21,1	13,2	19,6	15,7	23,3	23,8	19,5	19,1	18,4
19,3	15,4	19,1	16,8	17,8	14,7	16,5	21,3	21,9	21,9	20,4
10,1	16,8	17,2	16,2	19,2	20,2	20,8	18,6	18,1	20,5	19,4

Hãy xét xem chiều cao của cây rừng có tuân theo luật phân phối bình thường $N(m, \sigma)$ hay không?

Chú ý rằng ở đây các tham số m và σ của phân phối là chưa biết, do đó $l = 2$ và ước lượng $\bar{x}$ của m ; ước lượng S^2 của σ^2 được tính như sau:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{55} x_i = 17,84; S^2 = \frac{1}{n-1} \sum_{i=1}^{55} (x_i - \bar{x})^2 = 8,53$$

Vì rằng ở đây ký hiệu x_i được dùng để chỉ số liệu đo đạc nên "trị thực tế" ta dùng ký hiệu n_i thay vì dùng ký hiệu x_i như các ví dụ ở trên.

	Trị thực tế n_i	P_i	Trị lý thuyết np_i	$\frac{(n_i - np_i)^2}{np_i}$
$(-\infty) - 12$	2 } 6	0,0228	1,254	5,274
12 - 14		0,0731	4,020	
14 - 16	8	0,1686	9,273	0,175
16 - 18	12	0,2576	14,168	0,332
18 - 20	16	0,2484	13,662	0,400
20 - 22	10 } 13	0,1519	8,354	12,633
22 - $(+\infty)$		0,0778	4,279	
Tổng	55	1	55	$1,018 = \chi^2(2);$ $P > 0,5$

$$H_0 : m = 17,84, \sigma^2 = 8,53; \min[np_i] = 5,274 > 5$$

$\chi^2 = 1,018$ với $K = r - l - 1 = 5 - 2 - 1 = 2$ độ tự do; $P > 0,5$. Vậy giả thuyết H_0 được chấp nhận, nghĩa là chiều cao của cây rừng tuân theo luật phân phối bình thường $N(17,84; \sqrt{8,53})$.

Chú ý rằng ở đây các trị xác suất P_i được tính từ hàm phân phối xác suất

$$\Phi(x) = \frac{1}{2\pi} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \text{ như sau:}$$

$$P_1 = \Phi\left(\frac{12-17,84}{\sqrt{8,53}}\right) - \Phi(-\infty) = 0,0228$$

$$P_2 = \Phi\left(\frac{14-17,84}{\sqrt{8,53}}\right) - \Phi\left(\frac{12-17,84}{\sqrt{8,53}}\right) = 0,0731$$

$$P_7 = \Phi(+\infty) - \Phi\left(\frac{22-17,84}{\sqrt{8,53}}\right) = 0,0778$$

2.2.2.3. Xác định khoảng tin tưởng của biến lượng

Như đã trình bày ở trên là:

$$\chi^2(n-1) = \frac{(n-1)S^2}{\sigma^2}$$

Do đó, khoảng tin tưởng ở mức xác suất sai lầm $P = 0,05$ (2 đuôi) ứng với $\alpha = 0,025$ (1 đuôi) đối với phân phối χ^2 sẽ như sau:

$$\chi_{0,975}^2(n-1) < \chi^2(n-1) < \chi_{0,025}^2(n-1)$$

$$\Rightarrow \chi_{0,975}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{0,025}^2(n-1)$$

$$\Rightarrow \frac{(n-1)S^2}{\chi_{0,025}^2(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi_{0,975}^2(n-1)}$$

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\chi_{0,025}^2(n-1)} < \sigma^2 < \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\chi_{0,975}^2(n-1)}$$

Ví dụ 2.(17): Nồng độ vitamin C của 17 mẫu nước ép cà chua (mg/100g) như sau: 16, 22, 21, 20, 23, 21, 19, 15, 13, 23, 17, 20, 29, 18, 22, 16, 25. Hãy xác định khoảng tin tưởng của biến lượng.

Với $n = 17$, $\sum x_i = 340$, $\sum (x_i - \bar{x})^2 = 254$, $S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = 15,88$,

với $\alpha = 0,025$, $\chi_{0,025}^2(17-1) = 28,8$ và $\chi_{0,975}^2(17-1) = 6,91$ thì khoảng tin tưởng của biến lượng sẽ như sau:

$$\frac{254}{28,8} < \sigma^2 < \frac{254}{6,91}$$

$\Rightarrow 8,82 < \sigma^2 < 36,76$ hay $2,97 < \sigma < 6,06$. Chú ý rằng, vì phân phối của S là phân phối đối xứng nên trị S = 3,98 không nằm ở trung tâm của khoảng tin tưởng.

Việc xác định khoảng tin tưởng của biến lượng có một ý nghĩa thực tế là thẩm tra lại xem một giá trị nào đó của biến lượng của quần thể đã được xác lập từ trước có còn phù hợp nữa hay không. Nếu giá trị của biến lượng thuộc về khoảng tin tưởng là còn phù hợp, còn nếu nằm ngoài khoảng tin tưởng là không còn phù hợp nữa.

2.3. PHÂN PHỐI STUDENT T(K)

2.3.1. Đặc điểm

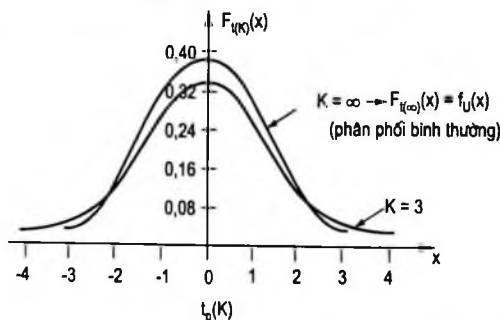
Nếu như có hai đại lượng ngẫu nhiên độc lập là U có phân phối bình thường dạng $N(0,1)$ và $\chi^2(K)$ có phân phối χ^2 với K độ tự do thì đại lượng ngẫu nhiên $\frac{U}{\sqrt{\frac{\chi^2(K)}{K}}}$ sẽ có phân phối Student với K độ tự do, tức là:

$$t(K) = \frac{U}{\sqrt{\frac{\chi^2(K)}{K}}}$$

Hàm mật độ của phân phối Student $f_{t(K)}(x)$ như sau:

$$f_{t(K)}(x) = \frac{G\left(\frac{K+1}{2}\right)}{G\left(\frac{K}{2}\right)\sqrt{\pi K}} \cdot \left(1 + \frac{x^2}{K}\right)^{-\frac{K+1}{2}}, \quad -\infty < x < +\infty$$

Đồ thị của hàm mật độ $f_{t(K)}(x)$ như sau (hình 2.6).



Hình 2.6. Đồ thị của hàm mật độ $f_{t(K)}(x)$ của phân phối Student với $K=3$ và $K=\infty$

phân phối Student $t(K)$ là một phân phối đối xứng có dạng như phân phối chuẩn. Phân phối Student $t(K)$ có trung bình $M[t(K)] = 0$ và biến lượng $[K] = \frac{K}{K-2}$, $K > 2$. Mối quan hệ giữa phân phối Student $t(K)$ và phân phối chuẩn dạng chuẩn $N(0,1)$ trong trường hợp K lớn được cho bởi công thức sau đây :

$$t_p(K) \approx U_p \left[\left(1 - \frac{1}{4K} \right)^2 - \frac{U_p^2}{2K} \right]^{\frac{1}{2}}$$

Do đó, trong trường hợp K lớn ($K > 30$) thì có thể coi $t_p(K) \approx U_p$ còn nếu $K \rightarrow \infty$ thì $t_p(\infty) = U_p$. Để minh họa chúng ta thử tính $t_{0,20}(4)$ theo công thức trên như sau:

$$t_{0,20}(40) \approx 1,282 \left[\left(1 - \frac{1}{4(40)} \right)^2 - \frac{1,282^2}{2(40)} \right]^{\frac{1}{2}} \approx 1,304$$

Ồ rõ ràng là giá trị này (1,304) không khác mấy so với giá trị đúng của nó $t_{0,20}(40) = 1,303$.

Chú ý rằng về mặt thực hành phân phối Student $t(K)$ thường được xác định như sau:

i) Nếu như các số liệu $x_1, x_2, \dots, x_n$ của mẫu được lấy từ quần thể có phân phối bình thường dạng $N(m, \sigma)$ với trung bình và biến lượng của mẫu được xác định như sau:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i; S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \text{ thì:}$$

$\bar{x}$ có phân phối bình thường dạng $N(m, \sqrt{\frac{\sigma^2}{n}})$.

$U = \frac{\bar{x} - m}{\sqrt{\frac{S^2}{n}}}$ có phân phối bình thường dạng $N(0, 1)$.

Còn đại lượng $\frac{\bar{x} - m}{\sqrt{\frac{S^2}{n}}}$ có phân phối Student với $n - 1$ độ tự do, nghĩa là:

$$t(n-1) = \frac{\bar{x} - m}{\sqrt{\frac{S^2}{n}}}$$

Phân phối Student $t(n-1)$ có trung bình là $M[t(n-1)] = 0$ và biến lượng là $V[t(n-1)] = \frac{(n-1)}{(n-3)}$.

Ở các phần trên chúng ta đã chứng minh phân phối của $\bar{x}$ và U . Ở đây chúng ta sẽ chứng minh phân phối của $t(n-1)$ như sau:

$$\frac{\bar{x} - m}{\sqrt{\frac{S^2}{n}}} = \frac{U \sqrt{\frac{\sigma^2}{n}}}{\sqrt{\frac{S^2}{n}}} = \frac{U}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{U}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} = t(n-1)$$

(ii) Nếu $\bar{x}_1$ và S_1^2 là trung bình và biến lượng của mẫu (n_1) được lấy từ quần thể có phân phối bình thường dạng $N(m_1, \sigma)$, còn $\bar{x}_2$ và S_2^2 là trung bình và biến lượng của mẫu (n_2) được lấy từ quần thể có phân phối bình thường dạng $N(m_2, \sigma)$, và S^2 là ước lượng của σ^2 được tính như sau:

$$S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1) + (n_2-1)} = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

thì :

$$- (\bar{x}_1 - \bar{x}_2) \text{ có phân phối bình thường dạng } N\left(m_1 - m_2, \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

$$- U = \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ có phân phối bình thường dạng chuẩn } N(0, 1).$$

$$\text{Còn đại lượng } \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ có phân phối Student với } n_1 + n_2 - 2$$

độ tự do nghĩa là:

$$t(n_1 + n_2 - 2) = \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Phân phối Student $t(n_1 + n_2 - 2)$ có trung bình là $M[t(n_1 + n_2 - 2)] = 0$ và
 phương sai là $V[t(n_1 + n_2 - 2)] = \frac{(n_1 + n_2 - 2)}{(n_1 + n_2 - 4)}$.

Ở các phần trên chúng ta đã chứng minh phân phối của $(\bar{x}_1 - \bar{x}_2)$ và U . Ở
 đây chúng ta sẽ chứng minh phân phối của $t(n_1 + n_2 - 2)$ như sau:

$$\frac{\bar{x}_1 - \bar{x}_2 - (m_1 - m_2)}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{U \cdot \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{U}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{U}{\sqrt{\frac{\chi^2(n_1 + n_2 - 2)}{n_1 + n_2 - 2}}} =$$

$t(n_1 + n_2 - 2)$.

(iii) Nếu $\bar{x}_1$ và S_1^2 là trung bình và biến lượng của mẫu (n_1) được lấy từ quần
 thể có phân phối bình thường dạng $N(m_1, \sigma_1^2)$ còn $\bar{x}_2$ và S_2^2 là trung bình và
 phương sai của mẫu (n_2) được lấy từ quần thể có phân phối bình thường dạng
 $N(m_2, \sigma_2^2)$ thì:

- $(\bar{x}_1 - \bar{x}_2)$ có phân phối bình thường dạng $N\left(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$

- $U = \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ phân phối bình thường dạng chuẩn $N(0, 1)$.

- Còn đại lượng $\frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)}}$ có phân phối Student với K độ tự

, nghĩa là:

$$t(K) = \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)}}$$

Ở đây độ tự do K được tính theo công thức sau đây:

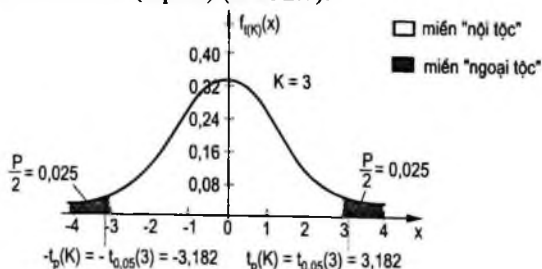
$$K = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2} \right)^2}{n_2 - 1}}$$

2.3.2. Ứng dụng

2.3.2.1. Trắc nghiệm t

Phân phối Student đóng một vai trò rất quan trọng trong thống kê học. Trắc nghiệm so sánh dựa trên căn bản của phân phối Student được gọi là "trắc nghiệm t".

Tương tự như trắc nghiệm U, để có thể thực hiện việc so sánh cần thiết lập các đại lượng cần so sánh nằm trong một biểu thức dưới dạng hiệu số và có phân phối Student. Dựa vào đặc điểm của phân phối Student là *trung bình của phân phối bằng zero* để rút ra kết luận rằng các đại lượng đem so sánh là như nhau hay khác nhau. Nếu biểu thức chứa các đại lượng cần so sánh là $t(K)$ nằm ở miền "nội tộc" của phân phối (miền ấn định chấp nhận giả thuyết tương đồng H_0) nghĩa là $t(K)$ nằm lân cận giá trị trung bình của phân phối, tức $t(K) \sim 0$, thì các đại lượng đem so sánh có thể coi là như nhau, trái lại, nếu $t(K)$ nằm ở miền "ngoại tộc" của phân phối (miền ấn định bác bỏ giả thuyết tương đồng H_0), nghĩa là $t(K)$ nằm xa giá trị trung bình của phân phối, tức $t(K) \neq 0$, thì các đại lượng đem so sánh là thực sự khác nhau. Như đã trình bày ở các phần trên, mức xác suất sai lầm $P = 0,05$ được dùng làm ranh giới phân chia giữa miền "ngoại tộc" (có $P \leq 0,05$) và miền "nội tộc" (có $P > 0,05$), và cũng như ở trắc nghiệm U, xác suất sai lầm P ở trắc nghiệm t cũng được tính từ 2 đuôi (2 phía) (hình 2.7).



Hình 2.7. Đồ thị minh họa miền "nội tộc" và miền "ngoại tộc" ở phân phối $t(K)$ với $K = 3$

Để minh họa chúng ta thử xét trường hợp:

$$t(n_1 + n_2 - 2) = \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \text{ có phân phối Student với } n_1 + n_2 - 2$$

2 độ tự do như sau:

Với giả thuyết tương đồng: $H_0 : m_1 = m_2$ thì $m_1 - m_2 = 0$.

$$\text{Do đó: } t(n_1 + n_2 - 2) = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Như vậy, nếu $t(n_1 + n_2 - 2)$ nằm ở miền “nội tộc” thì $t(n_1 + n_2 - 2) \sim 0 \Leftrightarrow \bar{x}_1 - \bar{x}_2 \sim 0 \Leftrightarrow \bar{x}_1 \sim \bar{x}_2$ (các đại lượng đem so sánh $\bar{x}_1$ và $\bar{x}_2$ được coi là như nhau), trái lại, nếu $t(n_1 + n_2 - 2)$ nằm ở miền “ngoại tộc” thì $t(n_1 + n_2 - 2) \neq 0 \Leftrightarrow \bar{x}_1 - \bar{x}_2 \neq 0 \Leftrightarrow \bar{x}_1 \neq \bar{x}_2$ (các đại lượng đem so sánh $\bar{x}_1$ và $\bar{x}_2$ được coi là khác nhau).

Sau đây là 3 ví dụ minh họa điển hình cho kiểu “trắc nghiệm t”. Các vấn đề khác liên quan tới các ví dụ này sẽ được trình bày chi tiết ở chương 3 “So sánh các nghiệm thức”.

Ví dụ 2.(18): So sánh hai nghiệm thức với số liệu “bắt cặp” (hay tạo “khối”): Để đánh giá ảnh hưởng của hormone A (testosterone) và C (dehydroandrosterone) tới sự phát triển của mào gà trống người ta đã thử nghiệm trên 22 con gà lấy ở 11 lứa khác nhau, kết quả như sau:

Hormone	Số gà	Khối lượng mào gà (mg)											Tổng	Trung bình
		Lứa (n = 11)												
		1	2	3	4	5	6	7	8	9	10	11		
A	11	57	120	101	127	119	117	104	73	53	68	118		
C	11	30	89	82	50	39	96	57	32	22	31	88		
Hiệu (x _i)		27	31	19	87	80	1	47	41	31	37	30	415	41

Chú ý là, vì rằng x_i có phân phối bình thường dạng $N(m, \sigma)$ với ước lượng của trung bình và biến lượng của phân phối như sau:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{27 + 31 + \dots + 30}{11} = \frac{451}{11} = 41$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] =$$

$$= \frac{1}{11-1} \left[27^2 + 31^2 + \dots + 30^2 - \frac{1}{11} (451)^2 \right] = 509$$

Do đó, $\bar{x}$ có phân phối bình thường dạng $N(m, \sqrt{\frac{\sigma^2}{n}})$ và $U = \frac{\bar{x} - m}{\sqrt{\frac{\sigma^2}{n}}}$ có

phân phối bình thường dạng chuẩn $N(0, 1)$, còn $t(n-1) = \frac{\bar{x} - m}{\sqrt{\frac{S^2}{n}}}$ có phân phối

Student với $n - 1$ độ tự do. Như vậy, với giả thuyết tương đồng $H_0 : m = 0$ thì: $t(n-1) = \frac{\bar{x}}{\sqrt{\frac{S^2}{n}}} \Rightarrow t(n-1) = \frac{41}{\sqrt{\frac{509}{11}}} = 6,03; P < 0,001$.

Vậy giả thuyết tương đồng H_0 bị bác bỏ, nghĩa là các nghiệm thức đem so sánh A và C thực sự khác nhau.

* *Chú ý:*

- Vì $m = m_1 = m_2$ nên giả thuyết tương đồng H_0 có thể đặt là $m = 0, m_1 = m_2 = 0$ hay $m_1 = m_2$.

- Nếu $t(n-1)$ nằm ở miền “nội tộc” (miền ấn định chấp nhận giả thuyết H_0) thì $t(n-1) \sim 0 \Rightarrow \bar{x} = x_1 - x_2 = \bar{x}_1 - \bar{x}_2 \sim 0 \Rightarrow \bar{x}_1 \sim \bar{x}_2$.

- Nếu $t(n-1)$ nằm ở miền “ngoại tộc” (miền ấn định bác bỏ giả thuyết H_0) thì $t(n-1) \neq 0 \Rightarrow \bar{x} = x_1 - x_2 = \bar{x}_1 - \bar{x}_2 \neq 0 \Rightarrow \bar{x}_1 \neq \bar{x}_2$.

Ví dụ 2.(19): So sánh hai nghiệm thức với số liệu không bắt “cặp” (hay không tạo “khối”): Để đánh giá tăng trưởng của cây được gieo ươm trên hai loại đất khác nhau (A và B) người ta tiến hành đo chiều cao một số cây con, kết quả như sau:

Loại đất	Số cây	Chiều cao (cm)	Tổng	Trung bình
A (x_1)	$n_1 = 12$	50; 41; 48; 60; 46; 60; 51; 42; 62; 54; 42; 46	$\Sigma x_1 = 602$	$\bar{x}_1 = 50,17$
B (x_2)	$n_2 = 10$	38; 40; 47; 51; 63; 50; 63; 57; 59; 51	$\Sigma x_2 = 519$	$\bar{x}_2 = 51,90$

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{12} (x_{1i} - \bar{x}_1)^2 = \frac{1}{n_1 - 1} \left[\sum_{i=1}^{12} x_{1i}^2 - \frac{1}{n_1} \left(\sum_{i=1}^{12} x_{1i} \right)^2 \right] = 55,06$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{10} (x_{2i} - \bar{x}_2)^2 = \frac{1}{n_2 - 1} \left[\sum_{i=1}^{10} x_{2i}^2 - \frac{1}{n_2} \left(\sum_{i=1}^{10} x_{2i} \right)^2 \right] = 76,32$$

Trước hết cần xác định xem σ_1 và σ_2 là như nhau hay khác nhau bằng “trắc nghiệm F” (sẽ được trình bày chi tiết ở phần sau) với giả thuyết tương đồng $H_0: \sigma_1 = \sigma_2$ như sau:

$$F(n_2 - 1, n_1 - 1) = \frac{S_2^2}{S_1^2} \Rightarrow F(9, 11) = \frac{76,32}{55,06} = 1,39; P > 0,1.$$

Vậy giả thuyết $H_0: \sigma_1 = \sigma_2$ được chấp nhận, nghĩa là $\sigma_1 = \sigma_2 = \sigma$.

Trong trường hợp này, chúng ta cần tiến hành xác định ước lượng S^2 của σ^2 như sau:

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(12 - 1)55,06 + (10 - 1)76,32}{(12 - 1) + (10 - 1)} = 64,627$$

Do đó, với giả thuyết tương đồng $H_0: m_1 = m_2$ thì:

$$t(n_1 + n_2 - 2) = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \Rightarrow t(20) = \frac{50,17 - 51,90}{\sqrt{64,627 \left(\frac{1}{12} + \frac{1}{10} \right)}} = -0,5;$$

$P > 0,5$.

Vậy giả thuyết tương đồng $H_0: m_1 = m_2$ được chấp nhận, nghĩa là các nghiệm thức đem so sánh A và B là như nhau.

Ví dụ 2.(20): So sánh hai nghiệm thức với số liệu không bắt “cấp” (hay không tạo “khối”): Tác dụng tăng lượng của cùng một chế độ dinh dưỡng như nhau trên hai nòi heo A và B như sau:

Nòi heo	Số heo	Tăng lượng (kg)	Tổng	Trung bình
A (x_1)	$n_1 = 5$	14,1; 10,1; 14,7; 13,7; 14,0	$\sum x_1 = 66,6$	$\bar{x}_1 = 13,32$
B (x_2)	$n_2 = 5$	14,0; 14,5; 13,7; 12,7; 14,1	$\sum x_2 = 69,0$	$\bar{x}_2 = 13,80$
$S_1^2 = 3,37; S_2 = 0,46$				

Trước hết cần xác định xem σ_1 và σ_2 là như nhau hay khác nhau bằng trắc nghiệm F với giả thuyết tương đồng $H_0: \sigma_1 = \sigma_2$ như sau :

$$F(n_1 - 1, n_2 - 2) = \frac{S_1^2}{S_2^2} \Rightarrow F(4, 4) = \frac{3,37}{0,46} = 7,33; P < 0,05$$

Vậy giả thuyết $H_0: \sigma_1 = \sigma_2$ bị bác bỏ, nghĩa là $\sigma_1 \neq \sigma_2$.

Trong trường hợp này, với giả thuyết tương đồng $H_0: m_1 = m_2$ thì:

$$t(K) = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{13,32 - 13,80}{\sqrt{\frac{3,37}{5} + \frac{0,46}{5}}} = -0,55$$

Với

$$K = \frac{\left(\frac{3,37}{5} + \frac{0,46}{5}\right)^2}{\frac{\left(\frac{3,37}{5}\right)^2}{4} + \frac{\left(\frac{0,46}{5}\right)^2}{4}} = 5$$

Vậy: $t(5) = -0,55$; $P > 0,5$, do đó, giả thuyết tương đồng $H_0: m_1 = m_2$ được chấp nhận, nghĩa là các nghiệm thức đem so sánh A và B là như nhau.

2.3.2.2. Khoảng tin tưởng và thuật so sánh dựa vào khoảng tin tưởng

Như đã trình bày ở trên, nếu đại lượng ngẫu nhiên X có phân phối bình thường dạng $N(m, \sigma)$ thì $\bar{x}$ có phân phối bình thường dạng $N(m, \sqrt{\frac{\sigma^2}{n}})$ và $U = \frac{\bar{x} - m}{\sqrt{\frac{\sigma^2}{n}}}$

có phân phối bình thường dạng chuẩn $N(0, 1)$, còn $t(n - 1) = \frac{\bar{x} - m}{\sqrt{\frac{S^2}{n}}}$ có phân phối

Student với $n - 1$ độ tự do, do đó, khoảng tin tưởng của trung bình của quần thể (m) ở mức xác suất sai lầm P sẽ như sau :

$$-t_p(n - 1) < t(n - 1) < t_p(n - 1) \Leftrightarrow -t_p(n - 1) < \frac{\bar{x} - m}{\sqrt{\frac{S^2}{n}}} < t_p(n - 1) \Leftrightarrow -t_p(n - 1)$$

$$\sqrt{\frac{S^2}{n}} < \bar{x} - m < t_p(n - 1) \sqrt{\frac{S^2}{n}} \Leftrightarrow \bar{x} - t_p(n - 1) \sqrt{\frac{S^2}{n}} < m < \bar{x} + t_p(n - 1) \sqrt{\frac{S^2}{n}}$$

$$\text{hay } m \in (\bar{x} - t_p(n-1) \cdot \sqrt{\frac{S^2}{n}}; \bar{x} + t_p(n-1) \cdot \sqrt{\frac{S^2}{n}}) \text{ hay } m = \bar{x} \pm t_p(n-1) \cdot \sqrt{\frac{S^2}{n}}$$

Lượng $t_p(n-1) \cdot \sqrt{\frac{S^2}{n}}$ được gọi là sai số cho phép (δ) ở mức xác suất sai lầm P .

Như thường lệ, nếu $P = 0,05$ thì được coi là mặc nhiên không cần phải nêu ra, còn nếu P ở các mức xác suất khác thì phải ghi rõ ra để tránh hiểu lầm.

Ví dụ 2(21): Nồng độ vitamin C của 17 mẫu nước ép cà chua (mg/100g) như sau: 16, 22, 21, 20, 23, 21, 19, 15, 13, 23, 17, 20, 29, 18, 22, 16, 25. Hãy xác định khoảng tin tưởng của trung bình.

Với: $n = 17$, $\sum x_i = 340$, $\bar{x} = 20$, $\sum (x_i - \bar{x})^2 = 254$,

$$S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = 15,88, \quad t_{0,05}(17-1) = 2,12 \text{ thì :}$$

$$m = \bar{x} \pm t_p(n-1) \cdot \sqrt{\frac{S^2}{n}} \Rightarrow m = 20 \pm 2,12 \cdot \sqrt{\frac{15,88}{17}}$$

Hay: $m = 20 \pm 2,05$.

Khoảng tin tưởng có thể được vận dụng để thực hiện việc so sánh các nghiệm thức. Để có thể tiến hành việc so sánh trước tiên cần thiết lập một biểu thức mà trong đó chứa các đại lượng cần so sánh nằm dưới dạng hiệu số và có phân phối Student, sau đó là xác định khoảng tin tưởng. Nếu khoảng tin tưởng bao gồm trị zero thì các đại lượng đem so sánh được coi là như nhau, trái lại, nếu khoảng tin tưởng không bao gồm trị zero thì các đại lượng đem so sánh được coi là khác nhau. Để minh họa chúng ta thử xét trường

hợp đại lượng $t(n_1 + n_2 - 2) = \frac{(x_1 - x_2) - (m_1 - m_2)}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ có phân phối Student

với $n_1 + n_2 - 2$ độ tự do, khoảng tin tưởng như sau:

$$(\bar{x}_1 - \bar{x}_2) - t_p(n_1 + n_2 - 2) \sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} < m_1 - m_2 < (\bar{x}_1 - \bar{x}_2) + t_p(n_1 + n_2 - 2) \sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Nếu khoảng tin tưởng bao gồm trị zero, nghĩa là $m_1 - m_2 \sim 0$ hay $m_1 \sim m_2$, trái lại, nếu khoảng tin tưởng không bao gồm trị zero, nghĩa là $m_1 - m_2 \neq 0$ hay $m_1 \neq m_2$.

Ví dụ 2.(22): Với số liệu ở ví dụ 2.(18) thì khoảng tin tưởng ở mức xác suất sai lầm $P = 0,05$ được xác định như sau:

$$\begin{aligned} \bar{x} - t_{0,05}(n-1) \sqrt{\frac{S^2}{n}} &< m < \bar{x} + t_{0,05}(n-1) \sqrt{\frac{S^2}{n}} \\ \Rightarrow 41 - 2,228 \cdot \sqrt{\frac{509}{11}} &< m < 41 + 2,228 \cdot \sqrt{\frac{509}{11}} \\ \Rightarrow 41 - 15,16 &< m < 41 + 15,16 \\ \Rightarrow 25,84 &< m < 56,16 \end{aligned}$$

Vì khoảng tin tưởng không bao gồm trị zero nên $m \neq 0 \Rightarrow m_1 - m_2 \neq 0 \Rightarrow m_1 \neq m_2$. Vậy các nghiệm thức đem so sánh A và C là khác nhau.

Ví dụ 2.(23): Với số liệu ở ví dụ 2.(19) thì khoảng tin tưởng ở mức xác suất sai lầm $P = 0,05$ được xác định như sau:

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) - t_p(n_1 + n_2 - 2) \sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} &< m_2 - m_1 < (\bar{x}_2 - \bar{x}_1) + t_p(n_1 \\ + n_2 - 2) \sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} &\Rightarrow (51,90 - 50,17) - 2,086 \sqrt{64,627 \cdot \left(\frac{1}{12} + \frac{1}{10} \right)} < \\ m_2 - m_1 < (51,90 - 50,17) + 2,086 \sqrt{64,627 \cdot \left(\frac{1}{12} + \frac{1}{10} \right)} &\Rightarrow 1,73 - 7,18 < \\ m_2 - m_1 < 1,73 + 7,18 &\Rightarrow -5,45 < m_2 - m_1 < 8,91. \end{aligned}$$

Vì khoảng tin tưởng bao gồm trị zero nên $m_2 - m_1 \sim 0 \Rightarrow m_2 \sim m_1$. Vậy các nghiệm thức đem so sánh A và B là như nhau.

Ví dụ 2.(24): Với số liệu ở ví dụ 2.(20) thì khoảng tin tưởng ở mức xác suất sai lầm $P = 0,05$ được xác định như sau:

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) - t_p(K) \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} &< m_2 - m_1 < (\bar{x}_2 - \bar{x}_1) + t_p(K) \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\ \Rightarrow (13,80 - 13,32) - 2,571 \cdot \sqrt{\frac{3,37}{5} + \frac{0,46}{5}} &< m_2 - m_1 < (13,80 - 13,32) \\ + 2,571 \cdot \sqrt{\frac{3,37}{5} + \frac{0,46}{5}} &\Rightarrow 0,48 - 2,25 < m_2 - m_1 < 0,48 + 2,25 \Rightarrow -1,77 < \\ m_2 - m_1 &< 2,73. \end{aligned}$$

Vì khoảng tin tưởng bao gồm trị zero nên $m_2 - m_1 \sim 0 \Rightarrow m_2 \sim m_1$. Vậy các nghiệm thức đem so sánh A và B là như nhau.

Ngoài ra, chúng ta cũng có thể dựa vào khoảng tin tưởng để so sánh nhiều nghiệm thức với nhau. Vấn đề này sẽ được trình bày chi tiết ở chương 3 "So sánh các nghiệm thức".

2.3.2.3. Ước định kích thước của mẫu

Như đã trình bày ở trên, khoảng tin tưởng của trung bình ở mức xác suất sai lầm P có thể được ghi dưới dạng $m = \bar{x} \pm t_p(n-1) \sqrt{\frac{S^2}{n}}$.

Lượng $t_p(n-1) \sqrt{\frac{S^2}{n}}$ được gọi là sai số cho phép (δ) ở mức xác suất sai lầm P. Vấn đề đặt ra ở đây là, với mong muốn đạt được một sai số cho phép không quá sai số ϵ quy định nào đó thì số liệu (kích thước của mẫu, n) cần thiết phải lấy sẽ là bao nhiêu. Để xác định n chúng ta cần tiến hành giải bất đẳng thức sau đây:

$$t_p(n-1) \sqrt{\frac{S^2}{n}} \leq \epsilon \Leftrightarrow n \geq \frac{t_p^2(n-1)S^2}{\epsilon^2}$$

Ví dụ 2.(25): Đo chiều cao của 50 cây rừng ta được kết quả sau đây: $n = 50$, $\bar{x} = 12,5$, $S^2 = 3,8$. Hãy xác định số cây cần phải đo để sai số cho phép của trung bình chiều cao của cây rừng không vượt quá 0,1 mét ở mức xác suất sai lầm $P = 0,05$.

Vì rằng $n \geq \frac{2^2 \cdot 3,8}{0,1^2} = 1520$, do đó, cần phải tiến hành đo tối thiểu là 1520 cây.

2.4. PHÂN PHỐI FISHER $F(K_1, K_2)$

2.4.1. Đặc điểm

Nếu như có hai đại lượng ngẫu nhiên độc lập là $\chi^2(K_1)$ có phân phối χ^2 với K_1 độ tự do và $\chi^2(K_2)$ có phân phối χ^2 với K_2 độ tự do thì đại lượng ngẫu

nhiên $\frac{\frac{\chi^2(K_1)}{K_1}}{\frac{\chi^2(K_2)}{K_2}}$ sẽ có phân phối Fisher với K_1 và K_2 độ tự do, tức là:

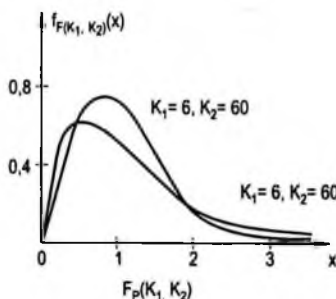
$$F(K_1, K_2) = \frac{\frac{\chi^2(K_1)}{K_1}}{\frac{\chi^2(K_2)}{K_2}}$$

Hàm mật độ $f_{F(K_1, K_2)}(x)$ của phân phối Fisher như sau:

$$f_{F(K_1, K_2)}(x) = \begin{cases} 0 & \text{nếu } x \leq 0 \\ \frac{G\left(\frac{K_1 + K_2}{2}\right)}{G\left(\frac{K_1}{2}\right)G\left(\frac{K_2}{2}\right)} \cdot \left(\frac{K_1}{K_2}\right)^{\left(\frac{K_1}{2}\right)} \cdot \frac{x^{\frac{K_1}{2}-1}}{\left(1 + \frac{K_1}{K_2}\right)^{\frac{K_1 + K_2}{2}}} & \text{nếu } x > 0 \end{cases}$$

Phân phối Fisher $F(K_1, K_2)$ có trung bình $M[F(K_1, K_2)] = \frac{K_2}{K_2 - 2}$, $K_2 > 2$.

Đồ thị của hàm mật độ $f_{F(K_1, K_2)}(x)$ của phân phối Fisher như sau (hình 2.8).



Hình 2.8. Đồ thị của hàm mật độ $f_{F(K_1, K_2)}(x)$ của phân phối Fisher với $K_1 = 6, K_2 = 6$ và $K_1 = 6, K_2 = 60$

Phân phối Fisher có đặc điểm đáng chú ý là:

$$F_{1-p}(K_1, K_2) = \frac{1}{F_p(K_1, K_2)}$$

Chẳng hạn:

$$F_{0,99}(3, 5) = \frac{1}{F_{0,01}(5, 3)} = \frac{1}{28,24} = 0,035$$

Vì rằng:

$$t(K) = \frac{U}{\sqrt{\frac{\chi^2(K)}{K}}} \Rightarrow t^2(K) = \frac{U^2}{\frac{\chi^2(K)}{K}} = \frac{\chi^2(1)}{\frac{\chi^2(K)}{K}} = \frac{1}{\frac{\chi^2(K)}{K}} = F(1, K)$$

Do đó, “trắc nghiệm t” với K độ tự do có thể được thay thế bởi “trắc nghiệm F” với 1 và K độ tự do và ngược lại nhờ mối quan hệ giữa phân phối Student và phân phối Fisher như sau:

$$t^2(K) = F(1, K) \quad \text{hay} \quad t(K) = \sqrt{F(1, K)}$$

Sau đây là một vài trị thông dụng $t_p(K)$ của phân phối Student và $F_p(1, K)$ của phân phối Fisher để minh họa cho mối quan hệ nêu trên:

P(*)	0,100	0,050	0,025	0,010	0,005
$F_p(1,5)$	4,06	6,61	10,01	16,26	22,78
$t_p(5)$	2,015	2,571	3,164	4,032	4,773
$t_p^2(5)$	4,06	6,61	10,01	16,26	22,78

K	6	10	15	19	24	30	40	60	120	∞
$F_{0,05}(1,K)$	5,99	4,96	4,54	4,38	4,26	4,17	4,08	4,00	3,92	3,84
$t_{0,05}(K)$	2,447	2,228	2,131	2,093	2,064	2,042	2,021	2,000	1,980	1,960
$t_{0,05}^2(K)$	5,99	4,96	4,54	4,38	4,26	4,17	4,08	4,00	3,92	3,84

(*) Trị xác suất sai lầm P ở phân phối Fisher được tính từ 1 đuôi, còn ở phân phối Student được tính từ 2 đuôi.

Vì rằng:

$$F(K_1, K_2) = \frac{\frac{\chi^2(K_1)}{K_1}}{\frac{\chi^2(K_2)}{K_2}} = \frac{\frac{\chi^2(K_1)}{K_1}}{\left[\frac{U}{t(K_2)} \right]^2}$$

Do đó:

$$f(K, \infty) = \frac{\frac{\chi^2(K)}{K}}{\left[\frac{U}{t(\infty)} \right]^2} = \frac{\frac{\chi^2(K)}{K}}{1^2} = \frac{\chi^2(K)}{K}$$

Nhờ đẳng thức $F(K, \infty) = \frac{\chi^2(K)}{K}$ hay $\chi^2(K) = K \cdot F(K, \infty)$ giúp ta hiểu

được mối quan hệ giữa “trắc nghiệm F” và “trắc nghiệm χ^2 ” cũng như giúp ta tính được trị gần đúng của F trong trường hợp K lớn, chẳng hạn:

$$F_{0,10}(4, 100) \approx \frac{\chi_{0,10}^2(4)}{4} = \frac{7,78}{4} = 1,945$$

Sau đây là một vài trị thông dụng $F_p(K, \infty)$ và $\chi_p^2(K)$ để minh họa cho mối quan hệ giữa phân phối Fisher và phân phối χ^2 nêu trên:

P(1 đuôi)	0,100	0,050	0,025	0,010	0,005
$F_p(5, \infty)$	1,85	2,21	2,57	3,02	3,35
$\chi_p^2(5)$	9,24	11,1	12,8	15,1	16,7
$\frac{\chi_p^2(5)}{5}$	1,85	2,21	2,57	3,02	3,35

K	6	10	15	20	24	30	40	60	120	∞
$F_{0,05}(K, \infty)$	2,10	1,83	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1
$\chi_{0,05}^2(K)$	12,6	18,3	25,0	31,4	36,4	43,8	55,8	79,1	146,4	∞
$\frac{\chi_{0,05}^2(K)}{K}$	2,10	1,83	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1

Chú ý rằng, từ đẳng thức $F(K, \infty) = \frac{\chi^2(K)}{K}$, nếu như $K = 1$ thì:

$F(1, \infty) = \chi^2(1)$, do đó, trắc nghiệm χ^2 với 1 độ tự do tương đương với trắc nghiệm F với 1 và ∞ độ tự do. Ngoài ra, như đã trình bày ở trên rằng

$\chi^2(1) = U^2$ và $U = t(\infty)$, do đó ta có được mối quan hệ sau đây giữa các phân phối:

$$F(1, \infty) = \chi^2(1) = U^2 = t^2(\infty)$$

Sau đây là một vài trị thông dụng $F_p(1, \infty)$, $\chi_p^2(1)$, U_p và $t_p(\infty)$ để minh họa cho mối quan hệ nêu trên giữa các phân phối:

P(*)	0.500	0.250	0.100	0.050	0.025	0.010	0.005
$F_p(1, \infty)$	0,455	1,32	2,71	3,84	5,02	6,63	7,88
χ_p^2	0,455	1,32	2,71	3,84	5,02	6,63	7,88
U_p	0,674	1,149	1,645	1,960	2,241	2,576	2,807
U_p^2	0,455	1,32	2,71	3,84	5,02	6,63	7,88
$t_p(\infty)$	0,674	1,149	1,645	1,960	2,241	2,576	2,807
$t_p^2(\infty)$	0,455		2,71	3,84	5,02	6,63	7,88

(*) Trị xác suất sai lầm P ở phân phối Fisher và phân phối χ^2 được tính từ 1 đuôi (1 phía), còn ở phân phối bình thường $N(0, 1)$ và phân phối Student được tính từ 2 đuôi (2 phía).

Ngoài ra, trong trường hợp K_1 và K_2 lớn thì phân phối Fisher có dạng gần như đối xứng, do đó, có thể được xác định thông qua phân phối bình thường dạng chuẩn $N(0, 1)$ như sau:

$$F_p(K_1, K_2) \approx \frac{K_2}{K_2 - 2} \cdot \left(1 + U_{2(1-p)} \sqrt{\frac{2(K_1 + K_2 - 2)}{K_1(K_2 - 4)}} \right) \text{ nếu } P \leq 0,5$$

$$F_p(K_1, K_2) \approx \frac{K_2}{K_2 - 2} \cdot \left(1 - U_{2(1-p)} \sqrt{\frac{2(K_1 + K_2 - 2)}{K_1(K_2 - 4)}} \right) \text{ nếu } P > 0,5$$

Chẳng hạn:

$$F_{0,05}(120, 60) \approx \frac{60}{60 - 2} \cdot \left(1 + 1,645 \sqrt{\frac{2(120 + 60 - 2)}{120(60 - 4)}} \right) \approx 1,43$$

(trị đúng là 1,47).

$$F_{0,95}(60, 120) \approx \frac{120}{120 - 2} \cdot \left(1 - 1,645 \sqrt{\frac{2(60 + 120 - 2)}{60(120 - 4)}} \right) \approx 0,639$$

(trị đúng được tính như sau: $F_{0,95}(60, 120) = \frac{1}{F_{0,05}(120, 60)} = \frac{1}{1,47} = 0,68$).

Chú ý rằng, về mặt thực hành phân phối Fisher thường được xác định như sau: Nếu mẫu có n_1 số liệu được lấy từ quần thể có phân phối bình thường dạng $N(m_1, \sigma_1)$ và mẫu có n_2 số liệu được lấy từ quần thể có phân phối bình

thường dạng $N(m_2, \sigma_2)$ thì đại lượng $\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$ sẽ có phân phối Fisher với $n_1 - 1$ và

$n_2 - 1$ độ tự do, tức là: $F(n_1 - 1, n_2 - 1) = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$

Thực vậy, vì rằng:

$$\chi^2(n_1 - 1) = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \Rightarrow \frac{S_1^2}{\sigma_1^2} = \frac{\chi^2(n_1 - 1)}{n_1 - 1}$$

$$\chi^2(n_2 - 1) = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \Rightarrow \frac{S_2^2}{\sigma_2^2} = \frac{\chi^2(n_2 - 1)}{n_2 - 1}$$

Do đó:
$$F(n_1 - 1, n_2 - 1) = \frac{\frac{\chi^2(n_1 - 1)}{n_1 - 1}}{\frac{\chi^2(n_2 - 1)}{n_2 - 1}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$$

2.4.2. Ứng dụng

2.4.2.1. Trắc nghiệm F

Vì rằng $F(n_1 - 1, n_2 - 1) = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$, do đó, nếu đặt giả thuyết tương đồng là

Ho: $\sigma_1^2 = \sigma_2^2 (= \sigma^2)$ thì:
$$F(n_1 - 1, n_2 - 1) = \frac{S_1^2}{S_2^2}$$

Chú ý rằng, phân phối Fisher $F(n_1 - 1, n_2 - 1)$ có trung bình $M[F(n_1 - 1, n_2 - 1)] = \frac{n_2 - 1}{n_2 - 3}$ nằm lân cận giá trị 1, do đó:

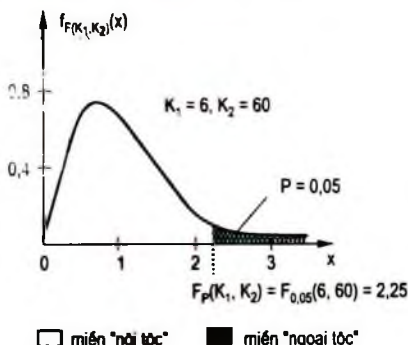
- Nếu trị $F(n_1 - 1, n_2 - 1)$ nằm ở miền "nội tộc" (miền ấn định chấp nhận giả thuyết H_0), nghĩa là $F(n_1 - 1, n_2 - 1)$ nằm lân cận giá trị trung bình của

phân phối, tức $F(n_1 - 1, n_2 - 1) \sim 1 \Rightarrow \frac{S_1^2}{S_2^2} \sim 1 \Rightarrow S_1^2 \sim S_2^2$ thì giả thiết H_0

được chấp nhận, nghĩa là $\sigma_1^2 = \sigma_2^2$.

- Nếu trị $F(n_1 - 1, n_2 - 1)$ nằm ở miền "ngoại tộc" (miền ấn định bác bỏ giả thuyết H_0), nghĩa là $F(n_1 - 1, n_2 - 1)$ nằm xa giá trị trung bình của phân phối, tức $F(n_1 - 1, n_2 - 1) \neq 1 \Rightarrow \frac{S_1^2}{S_2^2} \neq 1 \Rightarrow S_1^2 \neq S_2^2$, thì giả thuyết H_0 bị bác bỏ, nghĩa là $\sigma_1^2 \neq \sigma_2^2$.

Như thường lệ, mức xác suất sai lầm $P = 0,05$ được dùng làm ranh giới phân chia giữa miền "nội tộc" và miền "ngoại tộc". Các giá trị của F nằm ở miền "ngoại tộc" (bác bỏ giả thuyết H_0) có $P \leq 0,05$, còn các giá trị của F nằm ở miền "nội tộc" (chấp nhận giả thuyết H_0) có $P > 0,05$ (hình 2.9).



Hình 2.9. Đồ thị minh họa miền "nội tộc" và miền "ngoại tộc" của phân phối $F(K_1, K_2)$ với $K_1 = 6$ và $K_2 = 60$

Các trắc nghiệm dựa trên căn bản của phân phối Fisher được gọi là "trắc nghiệm F ". Trắc nghiệm F đóng một vai trò hết sức quan trọng trong thống kê học, đặc biệt là áp dụng trắc nghiệm F vào việc so sánh các nghiệm thức mà chúng ta sẽ thấy ở các phần sau. Ở đây chỉ đưa ra một ví dụ minh họa đơn giản về việc áp dụng trắc nghiệm F để so sánh các biến lượng trước khi tiến hành so sánh các nghiệm thức bằng trắc nghiệm t mà chúng ta đã gặp ở các phần trước.

Ví dụ 2.(26): Nghiệm thức thứ nhất có $n_1 = 15$ số liệu với trung bình $\bar{x}_1 = 187,6$ và biến lượng $S_1^2 = 1,45$. Nghiệm thức thứ hai có $n_2 = 14$ số liệu với trung bình

$\bar{x}_2 = 235,9$ và biến lượng $S_2^2 = 2,95$. Hãy so sánh biến lượng của chúng.

Với giả thuyết $H_0 : \sigma_1^2 = \sigma_2^2$ thì $F(n_2 - 1, n_1 - 1) = \frac{S_2^2}{S_1^2} \Rightarrow F(13, 14) = \frac{2,95}{1,45} = 2,03$; $P > 0,1$. Vậy giả thuyết H_0 được chấp nhận, nghĩa là $\sigma_1^2 = \sigma_2^2$.

* *Phụ chú:* Thực ra ở đây chúng ta có thể giải quyết vấn đề bằng cách dựa vào khoảng tin tưởng như sau: Vì rằng

$$F(n_1 - 1, n_2 - 1) = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}, \text{ do đó, khoảng tin tưởng ở mức xác suất sai lầm}$$

$P = 0,1$ (2 đuôi) ứng với $P = 0,05$ (1 đuôi) là:

$$F_{0,95}(n_1 - 1, n_2 - 1) < F(n_1 - 1, n_2 - 1) < F_{0,05}(n_1 - 1, n_2 - 1)$$

$$\Rightarrow F_{0,95}(n_1 - 1, n_2 - 1) < \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} < F_{0,05}(n_1 - 1, n_2 - 1)$$

$$\Rightarrow \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{0,05}(n_1 - 1, n_2 - 1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} \cdot F_{0,95}(n_1 - 1, n_2 - 1)$$

$$\Rightarrow \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{0,05}(n_1 - 1, n_2 - 1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} \cdot F_{0,05}(n_2 - 1, n_1 - 1)$$

$$\Rightarrow \frac{1,45}{2,95} \cdot \frac{1}{F_{0,05}(15 - 1, 14 - 1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1,45}{2,95} \cdot F_{0,05}(14 - 1, 15 - 1)$$

$$\Rightarrow \frac{1,45}{2,95} \cdot \frac{1}{2,56} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1,45}{2,95} \cdot 2,51$$

$$\Rightarrow 0,192 < \frac{\sigma_1^2}{\sigma_2^2} < 1,234.$$

Khoảng tin tưởng này bao gồm giá trị 1, nghĩa là $\frac{\sigma_1^2}{\sigma_2^2} \sim 1 \Rightarrow \sigma_1^2 \sim \sigma_2^2$. Chú

ý rằng, nếu như chúng ta tính khoảng tin tưởng của $\frac{\sigma_2^2}{\sigma_1^2}$ thì kết quả sẽ là

$0,811 < \frac{\sigma_2^2}{\sigma_1^2} < 5,208$, khoảng tin tưởng này cũng bao gồm giá trị 1.

2.4.2.2. Thuật so sánh dựa căn bản vào sự phân tích nguồn biến lượng

Như chúng ta đã biết, hầu hết các trắc nghiệm so sánh đều dựa trên căn bản là biến lượng của các quần thể thì bằng nhau, tức $\sigma_1^2 = \sigma_2^2 = \dots (= \sigma^2)$, còn trung bình của các quần thể là các đại lượng đem so sánh thì có thể bằng nhau hoặc khác nhau, do đó, trong thuật so sánh các nghiệm thức chúng ta chỉ cần đặt giả thuyết tương đồng H_0 rằng các trung bình là bằng nhau, tức $H_0: m_1 = m_2 = \dots (= m)$, sau đó tiến hành kiểm tra xem giả thuyết H_0 này là đúng hay sai (nhằm chấp nhận hay bác bỏ giả thuyết H_0) bằng cách dựa vào đặc điểm của các phân phối thường gặp trong thống kê học, đặc biệt là đặc điểm về giá trị trung bình của phân phối.

Thuật so sánh dựa căn bản vào sự phân tích nguồn biến lượng đóng một vai trò rất quan trọng trong việc so sánh các nghiệm thức, đặc biệt là so sánh nhiều (> 2) nghiệm thức với nhau. Sau đây là nội dung căn bản của thuật so sánh này:

Giả sử có l nghiệm thức (mẫu) độc lập được lấy từ l quần thể đều có phân phối bình thường với trung bình là $m_1, m_2, \dots, m_l$, và đều có biến lượng bằng nhau là σ^2 . Do đó ở đây để thực hiện việc so sánh các nghiệm thức với nhau chúng ta chỉ cần đặt giả thuyết tương đồng $H_0: m_1 = m_2 = \dots = m_l (= m)$ và dựa vào đặc điểm của phân phối Fisher có trung bình nằm lân cận giá trị 1 để kiểm tra giả thuyết H_0 này là đúng hay sai nhằm chấp nhận hay bác bỏ giả thuyết H_0 này. Tiến trình như sau: với X_{iK} là số liệu của nghiệm thức K ($i = 1, 2, \dots, n_K$; $K = 1, 2, \dots, l$) thì:

- Trung bình của nghiệm thức K là:

$$\bar{x}_K = \frac{1}{n_K} \cdot \sum_{i=1}^{n_K} x_{iK} = \frac{1}{n_K} \cdot x_K, \text{ với } x_K \text{ là tổng của nghiệm thức } K. \text{ Chú ý rằng}$$

trung bình của nghiệm thức $\bar{x}_K$ là ước lượng tốt nhất của m_K .

- Trung bình chung là:

$$\bar{x} = \frac{1}{n} \sum_{K=1}^I \sum_{i=1}^{n_K} x_{iK} = \frac{1}{n} \cdot x_{..}, \text{ với } x_{..} \text{ là tổng của các nghiệm thức, } x_{..} = \sum_{K=1}^I x_K,$$

và n là tổng số liệu của các nghiệm thức, $n = \sum_{K=1}^I n_K$. Chú ý rằng, trung bình chung $\bar{x}$ là ước lượng tốt nhất của m .

$$\text{Vì rằng: } (x_{iK} - \bar{x}) = [(\bar{x}_K - \bar{x}) + (x_{iK} - \bar{x}_K)]$$

$$\Rightarrow (x_{iK} - \bar{x})^2 = (\bar{x}_K - \bar{x})^2 + (x_{iK} - \bar{x}_K)^2 + 2(\bar{x}_K - \bar{x})(x_{iK} - \bar{x}_K)$$

$$\Rightarrow \sum_{K=1}^I \sum_{i=1}^{n_K} (x_{iK} - \bar{x})^2 = \sum_{K=1}^I n_K (\bar{x}_K - \bar{x})^2 + \sum_{K=1}^I \sum_{i=1}^{n_K} (x_{iK} - \bar{x}_K)^2 + \sum_{K=1}^I \sum_{i=1}^{n_K} 2(\bar{x}_K - \bar{x})(x_{iK} - \bar{x}_K) \text{ nhưng vì } \sum_{K=1}^I \sum_{i=1}^{n_K} 2(\bar{x}_K - \bar{x})(x_{iK} - \bar{x}_K) = 0,$$

do đó:

$$\boxed{\sum_{K=1}^I \sum_{i=1}^{n_K} (x_{iK} - \bar{x})^2 = \sum_{K=1}^I n_K (\bar{x}_K - \bar{x})^2 + \sum_{K=1}^I \sum_{i=1}^{n_K} (x_{iK} - \bar{x}_K)^2}$$

(Với $n - 1$ độ tự do) (Với $I - 1$ độ tự do) (Với $n - I$ độ tự do)

Đây là biểu thức căn bản phân tích nguồn biến lượng theo kiểu thu thập số liệu đơn giản nhất. Sự phân tích nguồn biến lượng theo những kiểu thu thập số liệu phức tạp hơn sẽ được trình bày chi tiết ở chương 3 "So sánh các nghiệm thức". Chú ý rằng:

$\sum_{K=1}^I \sum_{i=1}^{n_K} (x_{iK} - \bar{x})^2$ là tổng các bình phương của toàn thể số liệu lệch từ trung bình chung (total sum of squares), thường được ký hiệu là **Q** (tổng) hay **Q**.

$\sum_{K=1}^I n_K (\bar{x}_K - \bar{x})^2$ là tổng các bình phương của toàn thể đại diện nghiệm thức lệch từ trung bình chung (treatment sum of squares) thường được ký hiệu là **Q** (nghiệm thức) hay **Q_t**. **Q_t** biểu thị mức độ thăng giáng của các nghiệm thức.

Chú ý rằng $S_t^2 = \frac{Q_t}{(I-1)}$ là một ước lượng độc lập của σ^2 khi giả thuyết

tương đồng $H_0: m_1 = m_2 = \dots = m_I (= m)$ đúng.

$-\sum_{K=1}^l \sum_{i=1}^{n_K} (x_{iK} - \bar{x}_K)^2$ là tổng các bình phương của toàn thể số liệu trong

thí nghiệm lệch từ trung bình của thí nghiệm (error sum of squares), thường được ký hiệu là Q (sai lệch ngẫu nhiên) hay Q_e . Q_e biểu thị sự thăng giáng ngẫu nhiên của các số liệu. Chú ý rằng $S^2 = \frac{Q_e}{(n-l)}$ là ước lượng tốt

nhất σ^2 . Như chúng ta sẽ thấy ở các phần sau, S^2 chính là giá trị trung bình của các biến lượng S_K^2 của các thí nghiệm.

Bây giờ nếu đặt giả thuyết tương đồng là $H_0 : m_1 = m_2 = \dots m_l (= m)$ thì các đại lượng ngẫu nhiên độc lập $\frac{(l-1)S_t^2}{\sigma^2}$ và $\frac{(n-l)S^2}{\sigma^2}$ sẽ có phân phối χ^2 với $l-1$ và $n-l$ độ tự do, nghĩa là:

$$\chi^2(l-1) = \frac{(l-1)S_t^2}{\sigma^2} \Rightarrow \frac{\chi^2(l-1)}{(l-1)} = \frac{S_t^2}{\sigma^2}$$

$$\chi^2(n-l) = \frac{(n-l)S^2}{\sigma^2} \Rightarrow \frac{\chi^2(n-l)}{(n-l)} = \frac{S^2}{\sigma^2}$$

do đó, đại lượng: $F(l-1, n-l) = \frac{\frac{\chi^2(l-1)}{(l-1)}}{\frac{\chi^2(n-l)}{(n-l)}} = \frac{\frac{S_t^2}{\sigma^2}}{\frac{S^2}{\sigma^2}} = \frac{S_t^2}{S^2}$ sẽ có phân phối

Fisher với $l-1$ và $n-l$ độ tự do. Như thường lệ, nếu giá trị của $F(l-1, n-l)$ nằm ở miền “nội tộc” (có $P > 0,05$) thì giả thuyết H_0 được chấp nhận, nghĩa là $m_1 = m_2 = \dots = m_l (= m)$, các thí nghiệm đem so sánh được coi là như nhau. Trong trường hợp này $\bar{x}$ và S^2 là các ước lượng tốt nhất của m và σ^2 . Để nhận thấy là vì $F(l-1, n-l)$ nằm lân cận giá trị trung bình của phân phối, do đó $F(l-1, n-l) \sim 1 \Rightarrow \frac{S_t^2}{S^2} \sim 1 \Rightarrow S_t^2 \sim S^2$. Còn nếu giá trị của $F(l-1,$

$n-l)$ nằm ở miền “ngoại tộc” (có $P \leq 0,05$) thì giả thuyết H_0 bị bác bỏ, nghĩa là có ít nhất là 2 trong số các trung bình $m_1, m_2, \dots, m_l$ không bằng nhau, khi đó chúng ta lại phải tiến hành thêm một trắc nghiệm khác nữa để xác định xem các thí nghiệm khác nhau là các thí nghiệm thức nào. Để nhận thấy là, trong trường hợp này vì $F(l-1, n-l)$ nằm xa giá trị trung bình của phân phối,

do đó $F(l-1, n-l) \neq 1 \Rightarrow \frac{S_1^2}{S^2} \neq 1 \Rightarrow S_1^2 \neq S^2$. Để dễ hình dung, trong trường

hợp giả thuyết H_0 được chấp nhận thì sự thăng giáng của các nghiệm thức $(\bar{x}_K)$ chỉ là sự thăng giáng ngẫu nhiên, trái lại, nếu giả thuyết H_0 bị bác bỏ thì các nghiệm thức $(\bar{x}_K)$ thực sự khác nhau. Ngoài ra, từ biểu thức phân tích nguồn biến lượng $Q = Q_t + Q_e$ cũng giúp ta hình dung được rằng, khi sự biến động giữa các nghiệm thức càng lớn, nghĩa là Q_t càng lớn, thì giá trị F càng lớn nên khả năng bác bỏ giả thuyết H_0 càng cao và ngược lại, khi sự biến động giữa các nghiệm thức không đáng kể, thì Q_t sẽ nhỏ nên giá trị F cũng nhỏ theo, do đó khả năng chấp nhận giả thuyết H_0 càng lớn.

Ví dụ 2.(27): Có 3 nhóm sinh viên của 3 quốc gia theo học tiếng Việt. Qua kỳ thi cho thấy số lỗi ở bài làm của các sinh viên đó như sau :

Nhóm K	Số sinh viên n_K	Số lỗi ở mỗi bài làm x_{iK}	Tổng x_K	Trung bình $\bar{x}_K$
1	7	1; 3; 2; 0; 1; 2; 1	10	1,43
2	5	2; 3; 2; 1; 4	12	2,4
3	3	4; 5; 3	12	4,0

Hỏi khả năng học tiếng Việt của 3 nhóm sinh viên nói trên có khác nhau không?

Ở đây $l = 3, n = 7 + 5 + 3 = 15, H_0 : m_1 = m_2 = m_3$

$$\text{Với } x_{..} = \sum_{K=1}^l \sum_{i=1}^{n_K} x_{iK} = \sum_{K=1}^l x_K = 10 + 12 + 12 = 34$$

$$\sum_{K=1}^l \sum_{i=1}^{n_K} x_{iK}^2 = 1^2 + 3^2 + \dots + 3^2 = 104$$

Các đại lượng Q, Q_t và Q_e được tính nhanh như sau:

$$Q = \sum_{K=1}^l \sum_{i=1}^{n_K} x_{iK}^2 - \frac{1}{n} x_{..}^2 = 104 - \frac{1}{15} \cdot 34^2 = 26,93;$$

với $n - 1 = 15 - 1 = 14$ độ tự do.

$$Q_t = \sum_{K=1}^l \frac{1}{n_K} x_K^2 - \frac{1}{n} x_{..}^2 = \left(\frac{10^2}{7} + \frac{12^2}{5} + \frac{12^2}{3} \right) - \frac{34^2}{15} = 14,02;$$

với $l - 1 = 3 - 1 = 2$ độ tự do.

$$Q_e = Q - Q_1 = 26,93 - 14,02 = 12,91; \text{ với } n - l = 15 - 3 = 12 \text{ độ tự do.}$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Nghiem thức	2	14,02	
Sai lệch ngẫu nhiên	12	12,91	$S^2 = \frac{12,91}{12} = 1,08$
Tổng	14	26,93	
$F(2,12) = \frac{\frac{14,02}{2}}{\frac{12,91}{12}} = 6,52; P < 0,025$			

Vậy giả thuyết H_0 bị bác bỏ. nghĩa là khả năng học tiếng Việt của 3 nhóm sinh viên là không giống nhau. Do đó, chúng ta cần phân tích để xác định các nhóm khác nhau thực sự là những nhóm nào. Về vấn đề này có nhiều kiểu trắc nghiệm so sánh khác nhau và sẽ được giới thiệu đầy đủ ở chương 3, ở đây sẽ trình bày kiểu "trắc nghiệm Scheffe" (Scheffe's multiple F test) như sau :

Đại lượng $C = \sum_{K=1}^l \lambda_K m_K$, với $K = 1, 2, \dots, l$, được coi là một so sánh nếu

như $\sum_{K=1}^l \lambda_K = 0$. Ước lượng của C là $\bar{C} = \sum_{K=1}^l \lambda_K \bar{x}_K$, và ước lượng của biến

lượng của $\bar{C}$ là:

$$S_C^2 = V[\bar{C}] = V\left[\sum_{K=1}^l \lambda_K \bar{x}_K\right] = \sum_{K=1}^l V[\lambda_K \bar{x}_K] = \sum_{K=1}^l \lambda_K^2 \cdot \frac{S^2}{n_K} = S^2 \sum_{K=1}^l \frac{\lambda_K^2}{n_K}$$

Khoảng tin tưởng của C ở mức xác suất sai lầm P sẽ là:

$$C = \bar{C} \pm S_{\bar{C}} \sqrt{(l-1)(l-1, n-l)}$$

Như đã trình bày ở các phần trên về thuật so sánh dựa vào khoảng tin tưởng là, nếu khoảng tin tưởng của so sánh C bao gồm trị zero thì các nghiệm thức hay nhóm nghiệm thức đem so sánh là như nhau, trái lại, nếu khoảng tin tưởng của so sánh C không bao gồm trị zero thì các nghiệm thức hay nhóm nghiệm thức đem so sánh là khác nhau.

Để nhận thấy là, nếu chỉ có hai nghiệm thức, tức $l = 2$, thì $C = m_1 - m_2$,

$$\bar{C} = \bar{x}_1 - \bar{x}_2, \quad S_C^2 = S^2 \left[\frac{1^2}{n_1} + \frac{(-1)^2}{n_2} \right] = S^2 \left[\frac{1}{n_1} + \frac{1}{n_2} \right], \text{ và để ý là: } t_p(n-l) = \sqrt{F_p(l, n-l)}, \text{ thì: } (m_1 - m_2) = (\bar{x}_1 - \bar{x}_2) \pm t_p(n-l) \cdot \sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

Đây là khoảng tin tưởng của $(m_1 - m_2)$ mà chúng ta đã gặp ở phần áp dụng khoảng tin tưởng để so sánh hai nghiệm thức với nhau ở phần trình bày về phân phối Student.

Bây giờ chúng ta áp dụng kiểu trắc nghiệm này cho trường hợp của ví dụ trên ở mức xác suất sai lầm $P = 0,05$ như sau:

- So sánh nghiệm thức thứ nhất với nghiệm thức thứ hai (so sánh 1)

Đặt giả thuyết $H_0^{(1)}: m_1 = m_2$.

$$C_1 = m_1 - m_2; \lambda_1 = 1; \lambda_2 = -1, \lambda_3 = 0$$

$$\bar{C}_1 = \bar{x}_1 - \bar{x}_2 = 1,43 - 2,4 = 0,97$$

$$S_C^2 = 1,08 \cdot \left(\frac{1}{7} + \frac{1}{5} \right) = 0,37$$

Với $F_p(l-1, n-l) = F_{0,05}(2, 12) = 3,89$ thì:

$$\sqrt{(l-1)F_p(l-1, n-l)} = \sqrt{(3-1)3,89} = 2,79$$

Do đó: $(m_1 - m_2) = -0,97 \pm 1,697$ hay $(m_1 - m_2) \in (-2,667; 0,727)$

Vì khoảng tin tưởng của $(m_1 - m_2)$ là $(-2,667; 0,727)$ bao gồm trị zero, nghĩa là $m_1 - m_2 \sim 0 \Rightarrow m_1 \sim m_2$, do đó giả thuyết $H_0^{(1)}: m_1 = m_2$ được chấp nhận. Vậy khả năng học tiếng Việt của nhóm sinh viên thứ nhất và thứ hai là như nhau.

- So sánh nghiệm thức thứ nhất với nghiệm thức thứ ba (so sánh 2):

Đặt giả thuyết $H_0^{(2)}: m_1 = m_3$

$$C_2 = m_1 - m_3; \lambda_1 = 1; \lambda_2 = 0; \lambda_3 = -1$$

$$\bar{C}_2 = \bar{x}_1 - \bar{x}_3 = 1,43 - 4 = -2,57$$

$$S_{C_2}^2 = 1,08 \cdot \left(\frac{1}{7} + \frac{1}{3} \right) = 0,51$$

$(m_1 - m_3) = -2,57 \pm 1,992$ hay $(m_1 - m_3) \in (-4,562; -0,578)$

Vì khoảng tin tưởng của $(m_1 - m_3)$ là $(- 4,562; - 0,578)$ không bao gồm trị zero, nghĩa là $m_1 - m_3 \neq 0 \Rightarrow m_1 \neq m_3$, do đó giả thuyết $H_0^{(2)}: m_1 - m_3$ bị bác bỏ. Vậy khả năng học tiếng Việt của nhóm thứ nhất và thứ ba khác nhau.

- So sánh nghiệm thức thứ hai với nghiệm thức thứ ba (so sánh 3): Đặt giả thuyết $H_0^{(3)}: m_1 = m_3$

$$C_4 = m_2 - m_3; \lambda_1 = 0; \lambda_2 = 1; \lambda_3 = -1$$

$$\bar{C}_3 = \bar{x}_2 - \bar{x}_3 = 2,4 - 4 = -1,6$$

$$S_{\bar{C}_3}^2 = 1,08 \cdot \left(\frac{1}{5} + \frac{1}{3} \right) = 0,58$$

$(m_2 - m_3) = -1,6 \pm 2,125$ hay $(m_2 - m_3) \in (- 3,725; 0,525)$. Vì khoảng tin tưởng bao gồm zero, nghĩa là $m_2 - m_3 \sim 0 \Rightarrow m_2 \sim m_3$, do đó, giả thuyết $H_0^{(3)}$ được chấp nhận. Vậy khả năng học tiếng Việt của nhóm sinh viên thứ hai và thứ ba là như nhau.

Bây giờ giả sử rằng, sinh viên nhóm 1 và 2 là sinh viên của hai nước Châu Á, còn sinh viên nhóm ba là sinh viên của một nước Châu Âu. Như vậy, để đánh giá tình hình chúng ta cần có thêm một so sánh nữa là khả năng học tiếng Việt của sinh viên Châu Á và Châu Âu có khác nhau hay không như sau:

- So sánh sinh viên Châu Á và Châu Âu (so sánh 4):

Đặt giả thuyết $H_0^{(4)}: \frac{1}{2}(m_1 + m_2) = m_3$

$$C_3 = \frac{1}{2}(m_1 - m_2) - m_3; \lambda_1 = \frac{1}{2}; \lambda_2 = \frac{1}{2}; \lambda_3 = -1$$

$$\bar{C}_4 = \frac{1}{2}(\bar{x}_1 + \bar{x}_2) - \bar{x}_3 = \frac{1}{2}(1,43 + 2,4) - 4 = -2,085$$

$$S_{\bar{C}_4}^2 = 1,08 \cdot \left[\frac{\left(\frac{1}{2}\right)^2}{7} + \frac{\left(\frac{1}{2}\right)^2}{5} + \frac{1}{3} \right] = 0,45$$

$$\left[\frac{1}{2}(m_1 + m_2) - m_3 \right] = -2,085 \pm 1,872$$

Hay: $\left[\frac{1}{2}(m_1 + m_2) - m_3 \right] \in (-3,957; -0,213)$. Vì khoảng tin tưởng không

bao gồm trị zero, nghĩa là $\frac{1}{2}(m_1 + m_2) - m_3 \neq 0 \Rightarrow \frac{1}{2}(m_1 + m_2) \neq m_3$, do đó giả thuyết $H_0^{(4)}$ bị bác bỏ. Vậy khả năng học tiếng Việt của sinh viên Châu Á và Châu Âu khác nhau.

Chú ý rằng, trong tiến trình thực hiện việc so sánh các nghiệm thức chúng ta nên sắp xếp các nghiệm thức từ nhỏ đến lớn để có thể giảm bớt phần nào công việc tính toán. Ngoài ra, ngoài việc tiến hành so sánh đối chiếu các nghiệm thức với nhau, chúng ta còn có thể tiến hành so sánh đối chiếu các nhóm nghiệm thức với nhau, do đó, chúng ta cần cố gắng thiết lập “bộ so sánh” bao gồm các nghiệm thức và nhóm nghiệm thức sao cho có thể rút ra được nhiều thông tin nhất từ các kết quả nghiên cứu có được.

Chú ý rằng, căn bản của phép so sánh đối chiếu giữa hai nghiệm thức, giữa hai nhóm nghiệm thức hay giữa một nghiệm thức và một nhóm nghiệm thức là thiết lập khoảng tin tưởng của một so sánh (C), chẳng hạn như $C = \bar{C} \pm S_{\bar{C}} \sqrt{(l-1)F_p(l-1, n-l)}$. Ở đây đại lượng $S_{\bar{C}} \sqrt{(l-1)F_p(l-1, n-l)}$ không gọi là “sai số cho phép” ở mức xác suất sai lầm P mà được gọi là “sai biệt nhỏ nhất có ý nghĩa so sánh” (least significant difference) ở mức xác suất sai lầm P, thường được ký hiệu là $D_p^{(a)}$, hay “hạng sai biệt nhỏ nhất có ý nghĩa so sánh” (least significant range) ở mức xác suất sai lầm P, thường được ký hiệu là $R_p^{(a)}$, tùy thuộc vào kiểu so sánh đối chiếu là kiểu “trắc nghiệm không xếp hạng các nghiệm thức” (multiple test) hay là kiểu “trắc nghiệm có xếp hạng các nghiệm thức” (multiple range test). Nội dung của các kiểu trắc nghiệm này sẽ được trình bày chi tiết ở chương 3.

Từ ví dụ minh họa ở trên, chúng ta cũng dễ dàng thấy được là:

- Nếu $|\bar{C}| \geq D_p^{(a)}$ (hay $R_p^{(a)}$) thì khoảng tin tưởng của C không bao gồm trị zero, do đó, giả thuyết tương đồng H_0 bị bác bỏ ở mức xác suất sai lầm P, nghĩa là, các đại lượng đem so sánh (nghiệm thức hay nhóm nghiệm thức) là khác nhau ở mức xác suất sai lầm P.

- Nếu $|\bar{C}| < D_p^{(a)}$ (hay $R_p^{(a)}$) thì khoảng tin tưởng của C bao gồm giá trị zero, do đó giả thuyết tương đồng H_0 được chấp nhận ở mức xác suất sai lầm

P , nghĩa là các đại lượng đem so sánh (th nghiệm thức hay nhóm th nghiệm thức) là như nhau ở mức xác suất sai lầm P .

Chú ý rằng, về phương diện thống kê học, mọi kết luận đều phải chỉ rõ là ở mức xác suất sai lầm P nào, bởi vì ở mỗi mức xác suất sai lầm đòi hỏi P có thể có những kết luận rút ra không giống nhau. Mức xác suất sai lầm P càng nhỏ thì càng làm tăng sự chính xác và tính nghiêm ngặt ở các kết luận rút ra từ đó và ngược lại. Ba mức xác suất sai lầm P thường gặp trong các kết luận về phương diện thống kê học là 0,05; 0,01 và 0,001.

Bây giờ để hiểu rõ tại sao nói rằng biến lượng $S^2 = \frac{Q_e}{(n-l)}$ là ước lượng tốt nhất của biến lượng của quần thể σ^2 chúng ta cần chứng minh S^2 chính là giá trị trung bình của các biến lượng $S_1^2, S_2^2, \dots, S_l^2$, của các th nghiệm thức (mẫu). Thực vậy:

$$\begin{aligned}
 S_1^2 &= \frac{\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2}{n_1 - 1} \Rightarrow (n_1 - 1)S_1^2 = \sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 \\
 S_2^2 &= \frac{\sum_{i=1}^{n_2} (x_{i2} - \bar{x}_2)^2}{n_2 - 1} \Rightarrow (n_2 - 1)S_2^2 = \sum_{i=1}^{n_2} (x_{i2} - \bar{x}_2)^2 \\
 &\dots\dots\dots \\
 S_l^2 &= \frac{\sum_{i=1}^{n_l} (x_{il} - \bar{x}_l)^2}{n_l - 1} \Rightarrow (n_l - 1)S_l^2 = \sum_{i=1}^{n_l} (x_{il} - \bar{x}_l)^2 \\
 S^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \dots + (n_l - 1)S_l^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_l - 1)} \\
 &= \frac{\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 + \sum_{i=2}^{n_2} (x_{i2} - \bar{x}_2)^2 + \dots + \sum_{i=1}^{n_l} (x_{il} - \bar{x}_l)^2}{n_1 + n_2 + \dots + n_l - l} \\
 &= \frac{\sum_{K=1}^l \sum_{i=1}^{n_K} (x_{iK} - \bar{x}_K)^2}{n - l} = \frac{Q_e}{n - l}
 \end{aligned}$$

Chẳng hạn như ở ví dụ trên ta có:

$$(n_1 - 1)S_1^2 = \sum_{i=1}^7 (x_{i1} - \bar{x}_1)^2 = 1^2 + 3^2 + \dots + 1^2 - \frac{10^2}{7} = 5,71$$

$$(n_2 - 1)S_2^2 = \sum_{i=1}^5 (x_{i2} - \bar{x}_2)^2 = 2^2 + 3^2 + \dots + 4^2 - \frac{12^2}{5} = 5,2$$

$$(n_3 - 1)S_3^2 = \sum_{i=1}^3 (x_{i3} - \bar{x}_3)^2 = 4^2 + 5^2 + \dots + 3^2 - \frac{12^2}{3} = 2,0$$

$$S^2 = \frac{5,71 + 5,2 + 2}{15 - 3} = \frac{12,91}{12} = 1,08 \text{ (như trên)}$$

Chú ý rằng, biến lượng S^2 cũng như biến lượng của các nghiệm thức $S_1^2, S_2^2, \dots, S_l^2$ đều là biến lượng của các "thăng giáng ngẫu nhiên", trong khi đó biến lượng S_i^2 là biến lượng của các "thăng giáng nghiệm thức". Các thăng giáng nghiệm thức có thể là các thăng giáng ngẫu nhiên (trường hợp giả thuyết H_0 được chấp nhận), nhưng cũng có thể là các "thăng giáng thực sự" (trường hợp giả thuyết H_0 bị bác bỏ).

Chương 3

SO SÁNH CÁC NGHIỆM THỨC

3.1. KHÁI NIỆM VÀ CÁC VẤN ĐỀ LIÊN QUAN

3.1.1. Thuật so sánh trong thống kê học

Để so sánh đại lượng A với đại lượng B thường chúng ta thiết lập biểu thức chứa hiệu số (A - B) hay tỷ số $\frac{A}{B}$. Một khi điều kiện ràng buộc nào đó được thỏa mãn mà hiệu số (A - B) bằng zero hay coi như bằng zero hoặc tỷ số $\frac{A}{B}$ bằng 1 hay coi như bằng 1 thì A và B được coi là như nhau hay bằng

nhau, còn trong trường hợp ngược lại thì A và B được coi là khác nhau. Trong thống kê học người ta thường dựa vào các phân phối có trung bình bằng zero hay xấp xỉ bằng zero hoặc dựa vào các phân phối có trung bình bằng 1 hay xấp xỉ bằng 1 để thực hiện việc so sánh. Nếu hiệu số (A - B) nằm lân cận giá trị trung bình của phân phối (miền “nội tộc”, miền $P > 0,05$), tức $A - B \sim 0 \Rightarrow A \sim B$, thì A và B là như nhau (chấp nhận giả thuyết H_0), trái lại, nếu hiệu số (A - B) nằm xa giá trị trung bình của phân phối (miền “ngoại tộc”, miền $P \leq 0,05$), tức $A - B \neq 0 \Rightarrow A \neq B$, thì A và B là khác nhau (bác bỏ giả thuyết H_0). Tương tự, nếu tỷ số $\frac{A}{B}$ nằm lân cận giá trị trung bình của

phân phối (miền “nội tộc”, miền $P > 0,05$), tức $\frac{A}{B} \sim 1 \Rightarrow A \sim B$, thì A và B là

như nhau (chấp nhận giả thuyết H_0), trái lại, nếu tỷ số $\frac{A}{B}$ nằm xa giá trị

trung bình của phân phối (miền “ngoại tộc”, miền $P \leq 0,05$), tức $\frac{A}{B} \neq 1 \Rightarrow A$

$\neq B$, thì A và B là khác nhau (bác bỏ giả thuyết H_0). Phân phối bình thường dạng chuẩn $N(0,1)$ và phân phối student là các phân phối có trung bình bằng

zero, còn phân phối $\frac{\chi^2(K)}{K}$ là phân phối có trung bình bằng 1 và phân phối

Fisher là phân phối có trung bình nằm ở lân cận 1 nên thường được áp dụng

vào vấn đề so sánh các nghiệm thức. Ngoài ra, như đã trình bày ở các phần trên, chúng ta cũng có thể căn cứ vào các phân phối này để thiết lập khoảng tin tưởng của $(A - B)$ hay $\frac{A}{B}$ để thực hiện việc so sánh. Nếu khoảng tin

tưởng của $(A - B)$ ở mức xác suất sai lầm P bao gồm trị zero, nghĩa là $A - B \sim 0 \Rightarrow A \sim B$, thì A và B là như nhau (chấp nhận giả thuyết H_0) ở mức xác suất sai lầm P , trái lại, nếu khoảng tin tưởng của $(A - B)$ ở mức xác suất sai lầm P không bao gồm trị zero, nghĩa là $A - B \neq 0 \Rightarrow A \neq B$, thì A và B là khác nhau (bác bỏ giả thuyết H_0). Tương tự, nếu khoảng tin tưởng của $\frac{A}{B}$ ở

mức xác suất sai lầm P bao gồm giá trị 1, nghĩa là $\frac{A}{B} \sim 1 \Rightarrow A \sim B$, thì A và B là như nhau (chấp nhận giả thuyết H_0) ở mức xác suất sai lầm P , trái lại, nếu khoảng tin tưởng của $\frac{A}{B}$ ở mức xác suất sai lầm P không bao gồm giá

trị 1, nghĩa là $\frac{A}{B} \neq 1 \Rightarrow A \neq B$, thì A và B là khác nhau (bác bỏ giả thuyết

H_0). Thực ra, các thủ thuật so sánh trình bày ở trên về mặt hình thức thì trông có vẻ khác nhau nhưng xét về nội dung, bản chất của vấn đề thì lại hoàn toàn giống nhau. Chẳng hạn ở đây, thay vì trình bày như trên chúng ta có thể trình bày vấn đề theo cách khác như sau: nếu khoảng tin tưởng bao gồm giá trị trung bình của phân phối thì A và B là như nhau (chấp nhận giả thuyết H_0), còn nếu khoảng tin tưởng không bao gồm giá trị trung bình của phân phối thì A và B là khác nhau (bác bỏ giả thuyết H_0).

Thuật so sánh bằng cách lập hiệu số là phương thức rất phổ biến trong thống kê học dùng để so sánh các nghiệm thức với nhau. Chẳng hạn, để so sánh nghiệm thức A và nghiệm thức B chúng ta có thể lập hiệu số của trung bình $(\bar{x}_A - \bar{x}_B)$ hay lập hiệu số của tổng $(\sum x_{iA} - \sum x_{iB})$. Tương tự như vậy, để so sánh “nhóm nghiệm thức” gồm nghiệm thức A và nghiệm thức B với nghiệm thức kiểm chứng chúng ta có thể thiết lập hiệu số của trung bình $(\frac{\bar{x}_A + \bar{x}_B}{2} - \bar{x}_0)$ hay $[(\bar{x}_A + \bar{x}_B) - 2\bar{x}_0]$ hoặc thiết lập hiệu số của tổng

$$\left[\frac{\sum x_{iA} + \sum x_{iB}}{2} - \sum x_{i0} \right] \text{ hay } \left[(\sum x_{iA} + \sum x_{iB}) - 2\sum x_{i0} \right].$$
 Ngoài ra, để

so sánh “nhóm nghiệm thức” gồm nghiệm thức A và nghiệm thức B với “nhóm nghiệm thức” gồm nghiệm thức C , nghiệm thức D và nghiệm thức E

chúng ta có thể thiết lập hiệu số của trung bình $\left(\frac{\bar{x}_A + \bar{x}_B}{2} - \frac{\bar{x}_C + \bar{x}_D + \bar{x}_E}{3} \right)$

hay $[3(\bar{x}_A + \bar{x}_B) - 2(\bar{x}_C + \bar{x}_D + \bar{x}_E)]$ hoặc thiết lập hiệu số của tổng:

$$\left[\frac{\sum x_{iA} + \sum x_{iB}}{2} - \frac{\sum x_{iC} + \sum x_{iD} + \sum x_{iE}}{3} \right] \text{ hay } [3(\sum x_{iA} + \sum x_{iB}) - 2(\sum x_{iC} + \sum x_{iD} + \sum x_{iE})].$$

Cuối cùng cần chú ý rằng, tất cả các giá trị của xác suất P được trình bày trong quyển này là được tính từ 2 đuôi (2 phía) đối với phân phối bình thường $N(0,1)$ và phân phối Student và được tính từ 1 đuôi (1 phía) đối với phân phối χ^2 và phân phối Fisher.

3.1.2. Các phương pháp thu thập số liệu

Để so sánh cái này với cái kia chúng ta cần phải có số liệu. Số liệu có được từ kết quả của các cuộc thí nghiệm mà chúng ta đã tiến hành hay được lấy mẫu từ các quần thể có sẵn. Để cho khách quan thì số liệu ở mỗi nghiệm thức đem so sánh phải nhiều, có thể bằng nhau mà cũng có thể khác nhau. Số số liệu ở mỗi nghiệm thức có được từ các cuộc lặp lại của thí nghiệm hay lấy mẫu. Sau đây, là 3 phương pháp căn bản thu thập số liệu mà chúng ta cần phải nắm vững để có thể tiến hành phân tích biến lượng cho đúng.

3.1.2.1. Kiểu hoàn toàn ngẫu nhiên (completely random design)

Số liệu trong mỗi nghiệm thức cũng như số liệu ở các nghiệm thức khác nhau đều được lấy trong một hoàn cảnh (điều kiện) giống nhau. Trong trường hợp này, các số liệu trong mỗi nghiệm thức có thể hoán đổi vị trí cho nhau, số liệu đợt này có thể đem gán cho đợt kia mà không làm thay đổi kết quả phân tích biến lượng trong phép so sánh. Kiểu lấy số liệu như thế này thường được gọi là "kiểu hoàn toàn ngẫu nhiên" hay kiểu số liệu không tạo "khối".

3.1.2.2. Kiểu khối đầy đủ (randomized complete - block design)

Số liệu trong mỗi nghiệm thức được lấy từ những hoàn cảnh (điều kiện) khác nhau, nhưng số liệu của các nghiệm thức trong mỗi đợt thì được lấy trong một hoàn cảnh giống nhau.

Đợt 1 $\rightarrow$ $\boxed{T_1 \quad T_2 \dots T_l}$

(Khối 1) Hoàn cảnh H_1

Đợt 2 $\rightarrow$ $\boxed{T_1 \quad T_2 \dots T_l}$

(Khối 2) Hoàn cảnh H_2

.....
 Đợt $r \rightarrow \boxed{T_1 \quad T_2 \dots T_l}$

(Khối r) Hoàn cảnh H_r

Trong trường hợp này, số liệu của các nghiệm thức trong mỗi đợt kết thành “khối”, mỗi “khối” (mỗi đợt, mỗi lần lặp lại của nghiệm thức) như vậy thì số liệu của các nghiệm thức bắt “cặp” (nếu chỉ có 2 nghiệm thức), bắt “tam” (nếu có 3 nghiệm thức), bắt “tứ” (nếu có 4 nghiệm thức)... với nhau. Như vậy, trong trường hợp này trình tự các số liệu trong dãy ở mỗi nghiệm thức phải được giữ đúng, số liệu đợt này (hay khối này, lần lặp lại này) không thể đem gán cho đợt kia (hay khối kia, lần lặp lại kia) được, nghĩa là các số liệu trong dãy ở mỗi nghiệm thức không thể hoán đổi vị trí cho nhau được, bởi vì nếu trật tự các số liệu ở mỗi nghiệm thức bị đảo lộn thì sẽ có ảnh hưởng lớn tới kết quả phân tích biến lượng trong phép so sánh, dẫn đến sai lầm nghiêm trọng ở kết quả so sánh các nghiệm thức. Kiểu lấy số liệu theo “khối” như thế này thường được gọi là “kiểu khối đầy đủ” hay kiểu số liệu kết thành “khối”. Các nhà nghiên cứu thường bố trí các thí nghiệm của mình theo kiểu khối đầy đủ nhằm dễ theo dõi cũng như loại bỏ được các yếu tố ngoại lai tác động không đồng đều vào các nghiệm thức. Ngoài ra, trong quá trình thu thập số liệu ngoài thực địa, nếu các nghiệm thức nghiên cứu phân bố trong những hoàn cảnh (điều kiện) không đồng nhất thì nên thu thập số liệu theo kiểu khối đầy đủ để loại bỏ các yếu tố “hoàn cảnh” tác động không đồng đều vào các nghiệm thức nghiên cứu.

3.1.2.3. Kiểu bình phương Latin (Latin - square design)

Số liệu trong các nghiệm thức được lấy từ những đợt có những hoàn cảnh khác nhau (A_i) và trong mỗi đợt, số liệu của các nghiệm thức cũng được lấy trong những hoàn cảnh khác nhau (B_j).

	Hoàn cảnh $\rightarrow$	B_1	B_2	B_r
Đợt 1	A_1	T_1	T_2	T_r
Đợt 2	A_2	T_2	T_3	T_1
Đợt r	A_r	T_r	T_1	T_{r-1}

Trong trường hợp này, số liệu của các nghiệm thức vừa có thể kết thành “khối” theo hàng (đợt lặp lại, hoàn cảnh A_i) vừa có thể kết thành khối theo cột (hoàn cảnh B_j). Các “khối” là những hoàn cảnh giống nhau và đều chứa đầy đủ tất cả các nghiệm thức nghiên cứu dù đó là “khối hàng” hay là “khối cột”, do đó, số “khối hàng”, số “khối cột” và số nghiệm thức là bằng nhau và bằng số lần lặp lại của mỗi nghiệm thức. Kiểu lấy số liệu như thế này thường được gọi là “*kiểu bình phương Latin*” hay *kiểu số liệu kết thành “khối” cả hàng lẫn cột*. Thu thập số liệu theo kiểu bình phương Latin là nhằm loại bỏ các yếu tố “*hoàn cảnh đa phương*” tác động không đồng đều vào các nghiệm thức nghiên cứu. Đến đây chúng ta có thể mở rộng khái niệm và tổng quát hóa các phương pháp thu thập số liệu như sau: Chú ý rằng, “*hoàn cảnh đa phương*” ở “*kiểu bình phương Latin*” đúng ra là “*hoàn cảnh nhị phương*”. Như vậy một cách tổng quát là chúng ta có thể bố trí các nghiệm thức thí nghiệm trong “*hoàn cảnh n phương*” chứ không nhất thiết cứ phải là trong “*hoàn cảnh nhất phương*” như ở kiểu thu thập số liệu “*khối đầy đủ*” (kiểu số liệu kết thành “*khối*”), hay là trong “*hoàn cảnh nhị phương*” như ở kiểu thu thập số liệu “*bình phương Latin*” (kiểu số liệu kết thành “*khối*” cả hàng lẫn cột). Thậm chí chúng ta có thể mở rộng khái niệm này bằng cách dùng thuật ngữ “*hoàn cảnh vô phương*” để chỉ tính đồng nhất của hoàn cảnh mà trong đó các nghiệm thức nghiên cứu được thực hiện hay các số liệu của các nghiệm thức được thu thập từ đó, đây là trường hợp của kiểu thu thập số liệu “*hoàn toàn ngẫu nhiên*” (kiểu số liệu không tạo “*khối*”).

Tóm lại chúng ta có thể bố trí các thí nghiệm hoặc thu thập các số liệu của các nghiệm thức nghiên cứu trong điều kiện “*hoàn cảnh vô phương*” (kiểu số liệu không tạo “*khối*”), “*hoàn cảnh nhất phương*” (kiểu số liệu kết thành “*khối*”), “*hoàn cảnh nhị phương*” (kiểu số liệu kết thành “*khối*”, cả hàng lẫn cột), “*hoàn cảnh tam phương*”, “*hoàn cảnh tứ phương*”...

3.1.3. Thuật phân tích biến lượng

Phân tích biến lượng là một “công đoạn” quan trọng và có tính quyết định tới kết quả so sánh các nghiệm thức. Mọi lầm lẫn và sai sót trong phân tích biến lượng tất yếu dẫn tới các kết luận sai trái trong phép so sánh. Do đó, trước khi trình bày thuật so sánh các nghiệm thức nhất thiết phải nghiên cứu thật kỹ càng thuật phân tích biến lượng.

Như đã trình bày ở trên, sự phân tích nguồn biến lượng đặt căn bản trên phương pháp thu thập số liệu, do đó, tương ứng với 3 phương pháp căn bản thu thập số liệu là 3 phương thức phân tích nguồn biến lượng mà chúng ta sẽ lần lượt nghiên cứu sau đây:

3.1.3.1. Với số liệu không tạo “khối”

Trong những trường hợp mà số liệu của các nghiệm thức được lấy theo kiểu hoàn toàn ngẫu nhiên (kiểu số liệu không tạo “khối”) thì trật tự các số liệu trong dãy ở mỗi nghiệm thức không gây ảnh hưởng tới kết quả phân tích biến lượng, các số liệu ở mỗi nghiệm thức có thể hoán vị cho nhau. Trong trường hợp này chúng ta nói rằng số liệu của các nghiệm thức không tạo thành “khối”, nghĩa là các số liệu của các nghiệm thức không bắt “cặp” nhau (nếu chỉ có 2 nghiệm thức), không bắt “tam” nhau (nếu chỉ có 3 nghiệm thức), không bắt “tứ” nhau (nếu chỉ có 4 nghiệm thức)... Do đó, nguồn biến lượng chỉ bao gồm từ *sự thăng giáng của các nghiệm thức* ($\bar{x}_k$), với biến lượng là S_t^2 , và từ *sự thăng giáng ngẫu nhiên của các số liệu* (x_{ik}) đã lấy ở mỗi nghiệm thức, với biến lượng là S^2 . Chú ý rằng, sự thăng giáng của các nghiệm thức, thường được gọi là sự thăng giáng nghiệm thức, có thể là *sự thăng giáng ngẫu nhiên* (nếu giả thuyết H_0 được chấp nhận) nhưng cũng có thể là *sự thăng giáng thật sự* (nếu giả thuyết H_0 bị bác bỏ).

Như chúng ta đã thấy ở phần trước, về mặt toán học chúng ta đã chứng minh được biểu thức phân tích nguồn biến lượng trong trường hợp số liệu không tạo “khối” như sau:

$$\sum_{k=1}^l \sum_{i=1}^{n_k} (x_{ik} - \bar{x})^2 = \sum_{k=1}^l n_k (\bar{x}_k - \bar{x})^2 + \sum_{k=1}^l \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2$$

(với $n - 1$ độ tự do) (với $l - 1$ độ tự do) (với $n - l$ độ tự do)

Q Q_t Q_e

Điều mà chúng ta đã nhận định ở trên hoàn toàn phù hợp với biểu thức toán học phân tích nguồn biến lượng cho trường hợp số liệu không tạo “khối” này.

Từ biểu thức phân tích nguồn biến lượng chúng ta dễ dàng xác định được độ tự do của các thành tố Q , Q_t và Q_e như sau:

- Q được tính từ n số liệu nên độ tự do là $n - 1$
- Q_t được tính từ l nghiệm thức nên độ tự do là $l - 1$
- Q_e được tính từ l nhóm, mỗi nhóm có n_k số liệu nên có $n_k - 1$ độ tự do, do đó, độ tự do của Q_e sẽ là:

$$(n_1 - 1) + (n_2 - 1) + \dots + (n_l - 1) = (n_1 + n_2 + \dots + n_l) - l = n - l \text{ độ tự do.}$$

Chú ý rằng, với biểu thức phân tích nguồn biến lượng là $Q = Q_t + Q_e$ thì biểu thức của độ tự do tương ứng là $(n - 1) = (l - 1) + (n - l)$. Do đó, về

phương diện thực hành Q_e và độ tự do ($n - l$) của nó thường được tính một cách gián tiếp như sau: $Q_e = Q - Q_l$ và $n - l = (n - 1) - (l - 1)$.

Bây giờ chúng ta tìm hiểu về kỹ thuật tính toán các đại lượng Q , Q_l và Q_e - Q có thể được tính theo công thức định nghĩa là:

$$Q = \sum_{k=1}^l \sum_{i=1}^{n_k} (x_{ik} - \bar{x})^2 \text{ hoặc được tính nhanh bằng công thức sau đây:}$$

$$Q = \sum_{k=1}^l \sum_{i=1}^{n_k} x_{ik}^2 - \frac{1}{n} x^2 \dots; \text{ với } x_{..} = \sum_{k=1}^l \sum_{i=1}^{n_k} x_{ik}$$

- Q_l có thể được tính theo công thức định nghĩa là:

$$Q_l = \sum_{K=1}^l n_K (x_K - \bar{x})^2 \text{ hoặc được tính nhanh bằng công thức sau đây:}$$

$$Q_l = \sum_{k=1}^l \frac{1}{n_k} x_k^2 - \frac{1}{n} x^2 \text{ với } x_K = \sum_{i=1}^{n_k} x_i$$

- Q_e có thể được tính theo công thức định nghĩa là:

$$Q_e = \sum_{k=1}^l \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2 = \sum_{k=1}^l \left[\sum_{i=1}^{n_k} x_{ik}^2 - \frac{1}{n_k} x_k^2 \right]$$

hoặc được tính nhanh bằng công thức sau đây:

$$Q_e = Q - Q_l$$

Chú ý rằng, $S^2 = \frac{Q_e}{(n-l)}$ là ước lượng tốt nhất của biến lượng σ^2 của quần thể. Về bản chất, S^2 là giá trị trung bình của các biến lượng S_k^2 của các nghiệm thức được tính từ sự thăng giáng ngẫu nhiên của các số liệu nghiệm thức, nghĩa là:

$$S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \dots + (n_l-1)S_l^2}{(n_1-1) + (n_2-1) + \dots + (n_l-1)} = \frac{1}{n-l} \sum_{k=1}^l (n_k-1)S_k^2$$

$$\text{Vì rằng: } S_k^2 = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_{ik} - x_k)^2 \Rightarrow (n_k-1)S_k^2 = \sum_{i=1}^{n_k} (x_{ik} - x_k)^2$$

$$\Rightarrow \sum_{k=1}^l (n_k-1)S_k^2 = \sum_{k=1}^l \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2; \text{ do đó:}$$

$$S^2 = \frac{1}{n-l} \sum_{k=1}^l \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2 = \frac{1}{n-l} \cdot Q_e$$

Qua công thức này, một lần nữa ta kiểm chứng được rằng Q_e có độ tự do là $n - l$.

Sau đây là bảng *phân tích biến lượng* (analysis of variance) và trắc nghiệm F (F test), cho trường hợp số liệu không tạo “khối” mà cơ sở toán học của nó đã được trình bày đầy đủ ở chương 2.

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Nghiem thức	$l - 1$	Q_t	$S_t^2 = \frac{Q_t}{(l-1)}$
Sai lệch ngẫu nhiên	$n - l$	Q_e	$S^2 = \frac{Q_e}{(n-l)}$
Tổng	$n - 1$	Q	

$$F(l-1, n-l) = \frac{\frac{Q_t}{(l-1)}}{\frac{Q_e}{(n-l)}}; P$$

Để làm rõ nội dung của những khái niệm và kỹ thuật tính toán chúng ta thử xét tới trường hợp tổng quát nhất là số số liệu ở các nghiệm thức đều khác nhau sau đây :

Nghiem thức T_k	Số liệu	Tổng x_k	Số số liệu n_k	Trung bình $\bar{x}_k$
T_1	57,8; 56,2; 61,9; 54,4; 53,6; 56,4; 53,2.	393,5	7	$\bar{x}_1 = 56,21$
T_2	64,2; 58,7; 63,1; 62,5; 59,8; 59,2.	367,5	6	$\bar{x}_2 = 61,25$
T_3	62,5; 65,2; 67,6; 69,9; 69,4; 70,1; 67,8- 67,0; 68,5; 62,4.	670,4	10	$\bar{x}_3 = 67,04$
Tổng		$x_{..} = 1431,4$	$n = 23$	$\bar{x} = 62,23$

- Tính Q

+ Theo công thức định nghĩa:

$$Q = (57,8 - 62,23)^2 + (56,2 - 62,23)^2 + \dots + (62,4 - 62,23)^2 = 642,7$$

+ Tính nhanh:

$$Q = 57,8^2 + 56,2^2 + \dots + 62,4^2 - \frac{1431,4^2}{23} = 642,7$$

- Tính Q_i

+ Theo công thức định nghĩa:

$$Q_i = 7(56,21 - 62,23)^2 + 6(61,25 - 62,63)^2 + 10(67,04 - 62,63)^2 = 490,4$$

+ Tính nhanh:

$$Q_i = \frac{393,5^2}{7} + \frac{367,5^2}{6} + \frac{670,4^2}{10} - \frac{1431,4^2}{23} = 490,4$$

- Tính Q_e

+ Theo công thức định nghĩa:

Công thức 1:

$$Q_e = [(57,8 - 56,21)^2 + \dots + (53,2 - 56,21)^2] + [(64,2 - 61,25)^2 + \dots + (59,2 - 61,25)^2] + [(62,5 - 67,04)^2 + \dots + (62,4 - 67,04)^2] = 152,3$$

Công thức 2 (tính nhanh):

$$Q_e = \left(57,8^2 + \dots + 53,2^2 - \frac{393,5^2}{7} \right) + \left(64,2^2 + \dots + 59,2^2 - \frac{367,5^2}{6} \right) + \left(62,5^2 + \dots + 62,4^2 - \frac{670,4^2}{10} \right) = 152,3$$

Tính nhanh:

$$Q_e = 642,7 - 490,4 = 152,3$$

- Kiểm chứng lại kết quả tính toán: Dĩ nhiên khi cần phải kiểm tra lại kết quả tính toán thì Q_e chỉ được tính theo công thức định nghĩa (công thức 1 hoặc công thức 2). Kết quả tính toán là đúng nếu $Q_i + Q_e = Q$, là sai nếu $Q_i + Q_e \neq Q$. Ở đây rõ ràng là $490,4 + 152,3 = 642,7$ nên kết quả tính toán của chúng ta là đúng.

Kết quả phân tích biến lượng và trắc nghiệm F được trình bày ở bảng sau đây:

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Nghiem thức	2	490,4	$S_t^2 = 245,2$
Sai lệch ngẫu nhiên	20	152,3	$S^2 = 7,6$
Tổng	22	642,7	
$F(2, 20) = \frac{245,2}{7,6} = 32,26; P < 0,005$			

Vậy giả thuyết tương đồng $H_0: m_1 = m_2 = m_3$ bị bác bỏ, sự thăng giáng của các nghiệm thức là thực sự chứ không phải là ngẫu nhiên, nghĩa là tối thiểu cũng phải có ít nhất là 2 nghiệm thức khác nhau. Do đó, công việc tiếp theo của chúng ta là đối chiếu xem giữa các nghiệm thức cũng như giữa các nhóm nghiệm thức cái nào là giống nhau và cái nào là khác nhau. Vấn đề này sẽ được trình bày chi tiết ở các phần sau.

Trước khi chuyển sang mục khác, một lần nữa chúng ta trở lại vấn đề biến lượng S^2 và độ tự do của nó để thấy rõ rằng S^2 đích thực là biến lượng của các thăng giáng ngẫu nhiên bằng một khái niệm khác là giá trị “trung tâm” μ_K của các thăng giáng ngẫu nhiên của số liệu ở mỗi nghiệm thức. Sở dĩ vấn đề biến lượng của các thăng giáng ngẫu nhiên S^2 được lập đi lập lại và nhấn mạnh nhiều lần là vì tầm quan trọng của nó: S^2 là ước lượng tốt nhất, đúng đắn nhất của σ^2 (biến lượng của quần thể), nên được dùng làm đại diện cho σ^2 trong tất cả mọi phép tính liên quan.

Giả sử nghiệm thức K có n_K số liệu $x_{i,k}$ thăng giáng ngẫu nhiên quanh giá trị trung tâm μ_K . Trong trường hợp số liệu không tạo khối thì giá trị trung tâm của mỗi nghiệm thức sẽ là $\mu_K = \bar{x} + (\bar{x}_K - \bar{x}) = \bar{x}_K$. Như vậy trong trường hợp này giá trị trung tâm của các thăng giáng ngẫu nhiên trùng với giá trị trung bình của nghiệm thức ($\mu_K = \bar{x}_K$), do đó, Q_c và số độ tự do của nó sẽ được tính như sau:

$$\sum_{i=1}^{n_1} (x_{i1} - \mu_1)^2 = \sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 \quad \text{với } (n_1 - 1) \text{ độ tự do}$$

$$\sum_{i=1}^{n_2} (x_{i2} - \mu_2)^2 = \sum_{i=1}^{n_2} (x_{i2} - \bar{x}_2)^2 \quad \text{với } (n_2 - 1) \text{ độ tự do}$$

$$\sum_{i=1}^{n_l} (x_{il} - \mu_l)^2 = \sum_{i=1}^{n_l} (x_{il} - \bar{x}_l)^2 \quad \text{với } (n_l - 1) \text{ độ tự do}$$

$$Q_e = \sum_{k=1}^l \sum_{i=1}^{n_k} (x_{ik} - \mu_K)^2 = \sum_{k=1}^l \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2 \quad \text{với } (n_1 - l) \text{ độ tự do}$$

Rõ ràng là $S^2 = \frac{Q_e}{n-l}$ đích thực là biến lượng của các thăng giáng ngẫu nhiên. Chú ý rằng, vì $\mu_K \equiv \bar{x}_k$ với $K = 1, 2, \dots, l$ nên sẽ có l trung tâm, do đó độ tự do của Q_e sẽ là $(n - l)$.

3.1.3 2. Với số liệu kết thành “khối”

Trong trường hợp này số liệu của các nghiệm thức được lấy theo “khối”, mỗi khối là một “hoàn cảnh”, “hoàn cảnh” này có thể khác nhưng cũng có thể giống với “hoàn cảnh” kia. Nếu toàn bộ tất cả các “hoàn cảnh” ở các khối đều giống nhau thì đó là trường hợp của số liệu không tạo “khối” mà chúng ta vừa xem xét ở trên, còn nếu có thậm chí chỉ với 2 “hoàn cảnh” khác nhau thì đây là trường hợp của số liệu kết thành “khối”. Mỗi “khối” là một “cấp” (khối 2 nghiệm thức), một “bộ tam” (khối 3 nghiệm thức), một “bộ tứ” (khối 4 nghiệm thức)... số liệu của các nghiệm thức. Chính vì vậy mà trật tự của các số liệu trong dãy ở mỗi nghiệm thức phải được đảm bảo, không thể hoán vị cho nhau được, bởi lẽ trong từng khối các nghiệm thức hợp thành một “bộ sâu” nhờ tính “đồng cảnh ngộ” (cùng “hoàn cảnh” của chúng). Điều này cho phép các nghiệm thức nghiên cứu được lập lại theo khối trong những hoàn cảnh (điều kiện) khác nhau mà không sợ bị sai lầm khi so sánh, đánh giá kết quả nghiên cứu của các nghiệm thức. Như vậy rõ ràng là, trong trường hợp số liệu kết thành “khối” thì nguồn biến động sẽ bao gồm từ sự “thăng giáng nghiệm thức” (Q_t), từ sự “thăng giáng ngẫu nhiên” (Q_e) và từ sự thăng giáng của các khối nghiệm thức, thường gọi là sự “thăng giáng khối” hay “thăng giáng hoàn cảnh” (Q_b). Tương tự như sự thăng giáng “nghiệm thức”, sự “thăng giáng khối” (hoàn cảnh) có thể là sự thăng giáng thật sự (nếu giả thuyết tương đồng H_0 về hoàn cảnh bị bác bỏ) nhưng cũng có thể là sự thăng giáng ngẫu nhiên (nếu giả thuyết tương đồng H_0 về hoàn cảnh được chấp nhận). Từ nhận định trên, chúng ta có thể rút ra được biểu thức phân tích nguồn biến lượng trong trường hợp số liệu kết thành r “khối” như sau:

Q	$=$	Q_b	$+$	Q_t	$+$	Q_e
(Với $n - 1$ độ tự do)		(Với $r - 1$ độ tự do)		(Với $l - 1$ độ tự do)		(Với $n - r - l - 1$ độ tự do)

Bây giờ chúng ta tiến hành chứng minh biểu thức này như sau: Giả sử x_{ij} là số liệu của nghiệm thức i ($i = 1, 2, \dots, l$) nằm trong khối j ($j = 1, 2, \dots, r$). Dễ nhận thấy là $n = rl$ và $n_1 = n_2 = \dots = n_r = r$. Gọi $\bar{x}_i$ là trung bình của nghiệm thức i và $\bar{x}_j$ là trung bình của khối j .

Vì rằng:

$$\begin{aligned}(x_{ij} - \bar{x})^2 &= [(\bar{x}_j - \bar{x}) + (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_j + \bar{x})]^2 = (\bar{x}_j - \bar{x})^2 + (\bar{x}_i - \bar{x})^2 + \\ &+ (x_{ij} - \bar{x}_j + \bar{x}_i + \bar{x})^2 + 2(\bar{x}_j - \bar{x})(\bar{x}_i - \bar{x}) + 2(\bar{x}_j - \bar{x})(x_{ij} - \bar{x}_j - \bar{x}_i + \bar{x}) \\ &+ 2(\bar{x}_i - \bar{x})(x_{ij} - \bar{x}_j - \bar{x}_i + \bar{x})\end{aligned}$$

Do đó:

$\sum_{j=1}^r \sum_{i=1}^l (x_{ij} - \bar{x})^2 = \sum_{j=1}^r l(\bar{x}_j - \bar{x})^2 + \sum_{i=1}^l (x_{ij} - \bar{x})^2 + \sum_{j=1}^r \sum_{i=1}^l (x_{ij} - \bar{x}_j - \bar{x}_i + \bar{x})^2$
Q Q_b Q_t Q_e

Chú ý:

$$\begin{aligned}\sum_{j=1}^r \sum_{i=1}^l 2(\bar{x}_j - \bar{x})(\bar{x}_i - \bar{x}) &= \sum_{j=1}^r \sum_{i=1}^l 2(\bar{x}_j - \bar{x})(x_{ij} - \bar{x}_j - \bar{x}_i + \bar{x}) \\ &= \sum_{j=1}^r \sum_{i=1}^l 2(\bar{x}_i - \bar{x})(x_{ij} - \bar{x}_j - \bar{x}_i + \bar{x}) = 0\end{aligned}$$

Tương tự như trên, cách xác định số độ tự do của Q , Q_b , Q_t và Q_e như sau (hãy nhìn vào biểu thức phân tích biến lượng):

- Có $n = rl$ số liệu x_{ij} thẳng giáng quanh một trị trung tâm là $\bar{x}$ nên số độ tự do của Q là $n - 1$ hay $rl - 1$.

- Có r trị $\bar{x}_j$ thẳng giáng quanh một trị trung tâm $\bar{x}$ nên số độ tự do của Q_b là $r - 1$.

- Có l trị $\bar{x}_i$ thẳng giáng quanh trị trung tâm là $\bar{x}$ nên số độ tự do của Q_t là $l - 1$.

- Vì rằng $x_{ij} - \bar{x}_j - \bar{x}_i + \bar{x} = x_{ij} - (\bar{x}_j + \bar{x}_i - \bar{x})$, do đó, có n số liệu x_{ij} thẳng giáng ngẫu nhiên quanh $(r + l - 1)$ trị trung tâm là $(\bar{x}_j + \bar{x}_i - \bar{x})$ nên số độ tự do của Q_e là $[n - (r + l - 1)] = n - r - l + 1$ hay $(r - 1)(l - 1)$.

Thực ra, để xác định các trung tâm μ_{ij} của các thẳng giáng ngẫu nhiên

của số liệu x_{ij} chúng ta cần trở về với bản chất của vấn đề là $\mu_{ij} = \bar{x} + (\bar{x}_j - \bar{x}) + (\bar{x}_i - \bar{x}) = \bar{x}_j + \bar{x}_i - \bar{x}$, do đó, số lượng trung tâm của các thẳng giáng ngẫu nhiên sẽ là $r + l - 1$. Vì có n số liệu x_{ij} thẳng giáng ngẫu nhiên quanh $(r + l - 1)$ trị trung tâm μ_{ij} nên số độ tự do của Q_e sẽ là:

$$n - (r + l - 1) = n - r - l + 1 \text{ độ tự do.}$$

Các đại lượng Q và Q_t thì tính toán giống như ở phần trên, còn đại lượng Q_b có cách tính toán giống như cách tính toán đại lượng Q_l như sau:

- Theo công thức định nghĩa:

$$Q_b = \sum_{j=1}^r l(\bar{x}_j - \bar{x})^2$$

- Tính nhanh:

$$Q_b = \sum_{j=1}^r \frac{1}{l} x_j^2 - \frac{1}{n} x^2$$

Trong trường hợp số liệu kết thành khối, nếu căn cứ vào công thức định nghĩa để tính đại lượng Q_e thì rất là phức tạp nên chúng ta thường tính theo cách gián tiếp như sau: $Q_e = Q - (Q_b + Q_l)$.

Sau đây là bảng phân tích biến lượng và trắc nghiệm F cho trường hợp số liệu kết thành "khối":

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Khối (hoàn cảnh)	$r - 1$	Q_b	$S^2 = \frac{Q_e}{(n - r - l + 1)}$
Nghiem thức	$l - 1$	Q_l	
Sai lệch ngẫu nhiên	$n - r - l + 1$	Q_e	
Tổng	$n - 1$	Q	
$F(l-1, n-r-l-1) = \frac{\frac{Q_l}{(l-1)}}{\frac{Q_e}{(n-r-l-1)}}$			

Để làm rõ nội dung của những khái niệm và kỹ thuật tính toán chúng ta thử xét một ví dụ minh họa điển hình cho trường hợp số liệu kết thành khối là xem xét tác dụng tăng trọng của vitamin B12 đối với heo khác lứa sau đây:

Lượng $B_{12}(\mu\text{g})$	Tăng lượng (kg)			Tổng x_i	Trung bình $\bar{x}_i$
	Lứa 1	Lứa 2	Lứa 3		
0	1,12	1,08	1,01	3,21	1,07
5	1,26	1,21	1,19	3,66	1,22
10	1,29	1,23	1,23	3,75	1,25
15	1,38	1,27	1,22	3,87	1,29
Tổng, x_j	5,05	4,79	4,65	$x_{..} = 14,49$	$\bar{x} = 1,2075$

Chú ý rằng ở đây nếu như thí nghiệm được thực hiện trên heo cùng lứa (“hoàn cảnh” giống nhau) thì số liệu sẽ không kết thành khối, khi đó số liệu thí nghiệm sẽ được xử lý theo kiểu số liệu không tạo “khối”.

Trường hợp của ví dụ minh họa này là số liệu kết thành khối, mỗi khối là một “hoàn cảnh” (lứa heo) khác nhau. Thí nghiệm được thực hiện với 4 nghiệm thức ($l = 4$) được bố trí lặp lại trên 3 “hoàn cảnh” (lứa) khác nhau ($r = 3$). Kết quả tính toán như sau:

$$- Q = 1,12^2 + 1,08^2 + \dots + 1,22^2 - \frac{14,49^2}{12} = 0,107625$$

$$- Q_b = \frac{5,05^2}{4} + \frac{4,79^2}{4} + \frac{4,65^2}{4} - \frac{14,49^2}{12} = 0,0206$$

Với trung bình của khối (lứa) 1 là $\frac{5,05}{4} = 1,2625$, khối 2 là $\frac{4,79}{4} = 1,1975$

và khối 3 là $\frac{4,65}{4} = 1,1625$ chúng ta có thể tính Q_b theo công thức định nghĩa như sau:

$$Q_b = 4(1,2625 - 1,2075)^2 + 4(1,1975 - 1,2075)^2 + 4(1,1625 - 1,2075)^2 = 0,0206$$

$$Q_t = \frac{3,21^2}{3} + \frac{3,66^2}{3} + \frac{3,75^2}{3} + \frac{3,87^2}{3} - \frac{14,49^2}{12} = 0,083025$$

$$Q_e = 0,107625 - (0,0206 + 0,083025) = 0,004$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Lứa (khối)	2	0,0206	
Nghiệm thức	3	0,083025	
Sai lệch ngẫu nhiên	6	0,004	$S^2 = 0,00067$
Tổng	11	0,107625	...
$F(3, 6) = \frac{\frac{0,083025}{3}}{\frac{0,004}{6}} = 41,5125; P < 0,005$			

Vậy giả thuyết tương đồng $H_0: m_1 = m_2 = m_3 = m_4$ bị bác bỏ, sự thăng giáng của các nghiệm thức là một sự thăng giáng thực sự chứ không phải là một sự thăng giáng ngẫu nhiên. Nói một cách khác, ở đây phải có ít nhất là 2 nghiệm thức khác nhau thực sự trong số các nghiệm thức thí nghiệm. Việc so sánh chi tiết giữa hai nghiệm thức hoặc giữa các nhóm nghiệm thức sẽ được trình bày chi tiết ở phần sau.

3.1.3.3. Với số liệu kết thành “khối” cả hàng lẫn cột

Đây là trường hợp số liệu được lấy theo kiểu bình phương Latin. Số liệu của các nghiệm thức được lấy từ những đợt có những hoàn cảnh khác nhau (tạo “khối” theo hàng) và hơn nữa, trong mỗi đợt số liệu của các nghiệm thức cũng được lấy trong những hoàn cảnh khác nhau (tạo “khối” theo cột). Trong mỗi “khối hàng” hoặc mỗi “khối cột” đều hiện diện đầy đủ tất cả các nghiệm thức thí nghiệm. Như vậy, mỗi số liệu ở mỗi nghiệm thức làm “đại biểu” cho mỗi hoàn cảnh khác nhau, và nghiệm thức nào cũng hiện diện đầy đủ các “đại biểu” đại diện cho mỗi hoàn cảnh đó. Điều này cho phép các nghiệm thức thí nghiệm được lặp lại theo hoàn cảnh có tác động đa phương mà không sợ bị sai lầm khi so sánh, đánh giá kết quả nghiên cứu của các nghiệm thức. Ở kiểu bình phương Latin các nghiệm thức thí nghiệm được thực hiện trong hoàn cảnh có tác động nhị phương, dĩ nhiên, nếu cần chúng ta có thể bố trí các nghiệm thức thí nghiệm trong hoàn cảnh có tác động n phương. Khi môi trường tiến hành bố trí thí nghiệm hay thu thập số liệu không đồng nhất, có nhiều hoàn cảnh (điều kiện) khác nhau thì việc bố trí các nghiệm thức được lặp lại theo “hoàn cảnh n phương” là cần thiết nhằm loại bỏ yếu tố “hoàn cảnh n phương” tác động không đồng đều vào các nghiệm thức nghiên cứu.

Qua sự trình bày ở trên chúng ta cũng dễ dàng thấy được rằng, trong trường hợp số liệu kết thành “khối” cả hàng lẫn cột (kiểu bình phương Latin hay kiểu “hoàn cảnh nhị phương”) thì nguồn biến lượng sẽ bao gồm từ sự “thăng giáng nghiệm thức” (Q_t), từ sự “thăng giáng ngẫu nhiên” (Q_e), từ sự “thăng giáng khối hàng” hay sự “thăng giáng hoàn cảnh phương 1” (Q_r hay Q_1) và từ sự “thăng giáng khối cột” hay sự “thăng giáng hoàn cảnh phương 2” (Q_c hay Q_2). Tương tự như sự “thăng giáng nghiệm thức”, sự “thăng giáng khối hàng” cũng như sự “thăng giáng khối cột” có thể là sự thăng giáng ngẫu nhiên (nếu giả thuyết tương đồng H_0 về hoàn cảnh được chấp nhận) nhưng cũng có thể là sự thăng giáng thực sự (nếu giả thuyết tương đồng H_0 về hoàn cảnh bị bác bỏ). Từ nhận định trên chúng ta có thể rút ra được biểu thức phân tích nguồn biến động trong trường hợp số liệu kết thành “khối” cả hàng lẫn cột như sau:

$$Q = Q_r + Q_c + Q_t + Q_e$$

Tương tự như trên, về mặt toán học biểu thức này cũng được chứng minh như sau : giả sử x_{ij} là số liệu của nghiệm thức T nằm trong khối hàng thứ i và khối cột thứ j. Chú ý rằng trong trường hợp số liệu kết thành “khối” cả hàng lẫn cột thì số nghiệm thức, số lần lặp lại (số số liệu) của mỗi nghiệm thức, số khối hàng và số khối cột đều bằng nhau và bằng r, còn tổng số số liệu của các nghiệm thức $n = r^2$. Gọi $\bar{x}_i$ là trung bình của khối hàng i, $\bar{x}_t$ là trung bình của khối cột j và $\bar{x}_t$ là trung bình của nghiệm thức T. Vì rằng:

$$(x_{ij} - \bar{x})^2 = [(\bar{x}_i - \bar{x}) + (\bar{x}_j - \bar{x}) + (\bar{x}_t - \bar{x}) + (x_{ij} - \bar{x}_i - \bar{x}_j - \bar{x}_t + 2\bar{x})]^2$$

Do đó, triển khai tương tự như ở phần trước chúng ta sẽ có được kết quả là:

$$\begin{aligned} \sum_{j=1}^r \sum_{i=1}^r (x_{ij} - \bar{x})^2 &= \sum_{j=1}^r r(\bar{x}_j - \bar{x})^2 + \sum_{i=1}^r r(\bar{x}_i - \bar{x})^2 + \sum_{j=1}^r r(x_{tj} - \bar{x}_t)^2 + \\ &\quad + \sum_{j=1}^r \sum_{i=1}^r (x_{ij} - \bar{x}_i - \bar{x}_j - \bar{x}_t + 2\bar{x})^2 \end{aligned}$$

Q
 Q_r
 Q_c
 Q_t

Q_e

Căn cứ vào biểu thức phân tích nguồn biến lượng chúng ta có thể xác định số độ tự do của Q , Q_r , Q_c , Q_t và Q_e như sau:

- Có $n = r^2$ số liệu x_{ij} thăng giáng quanh một trị trung tâm là $\bar{x}$ nên số độ tự do của Q là $n - 1$ hay $r^2 - 1$.

- Có r trị $\bar{x}_i$ thăng giáng quanh một trị trung tâm là $\bar{x}$ nên số độ tự do của Q_r là $r - 1$.

- Có r trị $\bar{x}_j$ thăng giáng quanh một trị trung tâm là $\bar{x}$ nên số độ tự do của Q_c là $r - 1$.

- Có r trị $\bar{x}_t$ thăng giáng quanh một trị trung tâm là $\bar{x}$ nên số độ tự do của Q_t là $r - 1$.

Vì rằng: $x_{ij} - \bar{x}_i - \bar{x}_j - \bar{x}_t + 2\bar{x} = x_{ij} - (\bar{x}_i + \bar{x}_j + \bar{x}_t - 2\bar{x})$ do đó, có n số liệu x_{ij} thăng giáng ngẫu nhiên quanh $(r + r + r - 2) = (3r - 2)$ trị trung tâm là $(\bar{x}_i + \bar{x}_j + \bar{x}_t - 2\bar{x})$ nên số độ tự do của Q_e là $[n - (3r - 2)] = n - 3r + 2$ hay $(r - 1)(r - 2)$.

Thực ra, để xác định các trung tâm μ_{ijt} của các thăng giáng ngẫu nhiên của số liệu x_{ij} chúng ta cần trở về với bản chất của vấn đề là:

$\mu_{ijt} = \bar{x} + (\bar{x}_i - \bar{x}) + (\bar{x}_j - \bar{x}) + (\bar{x}_t - \bar{x}) = \bar{x}_i + \bar{x}_j + \bar{x}_t - 2\bar{x}$, do đó, số lượng trung tâm của các thăng giáng ngẫu nhiên sẽ là: $r + r + r - 2 = 3r - 2$. Vì có $n = r^2$ số liệu x_{ij} thăng giáng ngẫu nhiên quanh $(3r - 2)$ trị trung tâm μ_{ijt} nên độ tự do của Q_e sẽ là $n - (3r - 2) = n - 3r + 2$ hay $(r - 1)(r - 2)$ độ tự do.

Các đại lượng Q và Q_t thì tính toán giống như ở các phần trước, còn các đại lượng Q_r và Q_c thì cách tính toán giống như cách tính toán của Q_t như sau:

- Theo công thức định nghĩa:

$$+ Q_r = \sum_{i=1}^r r(\bar{x}_i - \bar{x})^2 \quad Q_c = \sum_{i=1}^r r(\bar{x}_j - \bar{x})^2$$

- Tính nhanh:

$$+ Q_r = \sum_{i=1}^r \frac{1}{r} \cdot x_i^2 - \frac{1}{n} \cdot x^2 \quad Q_c = \sum_{i=1}^r \frac{1}{r} \cdot x_j^2 - \frac{1}{n} \cdot x^2$$

Trong trường hợp số liệu kết thành khối có hàng lẫn cột, nếu căn cứ vào công thức định nghĩa để tính đại lượng Q_e thì rất là phức tạp nên chúng ta thường tính theo cách gián tiếp như sau:

$$Q_e = Q - (Q_r + Q_c + Q_t)$$

Sau đây là bảng phân tích biến lượng và trắc nghiệm F cho trường hợp số liệu kết thành “khối” cả hàng lẫn cột:

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Khối hàng	$r - 1$	Q_r	$S^2 = \frac{Q_e}{n - 3r + 2}$
Khối cột	$r - 1$	Q_c	
Nghiệm thức	$r - 1$	Q_t	
Sai lệch ngẫu nhiên	$n - 3r + 2$	Q_e	
Tổng	$n - 1$	Q	
$F(r - 1, n - 3r + 2) = \frac{\frac{Q_t}{r - 1}}{\frac{Q_e}{n - 3r + 2}}; P$			

Để làm rõ nội dung của những khái niệm và kỹ thuật tính toán chúng ta thử xét một ví dụ minh họa điển hình cho trường hợp số liệu kết thành khối cả hàng lẫn cột như sau: nghiên cứu tình hình số người đi coi nhạc kịch thuộc các lứa tuổi khác nhau : A (55 - 65), B (45 - 55), C (35 - 45), D (25 - 35) và E (15 - 25).

Hàng (tuần)	Cột (thứ)					Tổng x_i
	Bảy	Chủ nhật	Ba	Năm	Sáu	
1	B(257)	D(245)	E(182)	A(203)	C(231)	1118
2	E(230)	A(283)	B(252)	C(204)	D(271)	1240
3	A(279)	E(245)	C(280)	D(227)	B(266)	1297
4	C(287)	B(280)	D(246)	E(193)	A(334)	1340
5	D(202)	C(260)	A(250)	B(259)	E(338)	1309
Tổng, x_j	1255	1313	1210	1086	1440	$x = 6304$

	Nghiệm thức					Tổng
	A	B	C	D	E	
Tổng, x_i	1349	1314	1262	1191	1188	$x_{..} = 6304$
Trung bình, $\bar{x}_i$	269,8	262,8	252,4	238,2	237,6	$\bar{x} = 252,16$

Ở đây hoàn cảnh theo phương thứ 1 (khối hàng) là các tuần lễ khác nhau, còn hoàn cảnh theo phương thứ hai (khối cột) là các thứ khác nhau. Số

nghiệm thức, số số liệu (số lần lặp lại) ở mỗi nghiệm thức, số khối hàng và số khối cột đều bằng nhau và bằng 5. Tổng số các số liệu là $n = 5 \times 5 = 25$. Kết quả tính toán như sau:

$$- Q = 257^2 + 245^2 + \dots + 338^2 - \frac{6304^2}{25} = 36571,36$$

$$- Q_r = \frac{1118^2}{5} + \frac{1240^2}{5} + \dots + \frac{1309^2}{5} - \frac{6304^2}{25} = 6146,16$$

Với trung bình của khối hàng thứ nhất là $\frac{1118}{5} = 223,6$; thứ hai là $\frac{1240}{5} = 248,0$; thứ ba là $\frac{1297}{5} = 259,4$; thứ tư là $\frac{1340}{5} = 268,0$ và thứ năm là $\frac{1309}{5} = 261,8$ chúng ta có thể tính Q_r theo công thức định nghĩa như sau:

$$Q_r = 5(223,6 - 252,16)^2 + 5(248,0 - 252,16)^2 + \dots + 5(261,8 - 252,16)^2 = 6146,16$$

$$Q_c = \frac{1255^2}{5} + \frac{1313^2}{5} + \dots + \frac{1440^2}{5} - \frac{6304^2}{25} = 13601,36$$

Với trung bình của khối cột thứ nhất là $\frac{1255}{5} = 251,0$; thứ hai $\frac{1313}{5} = 262,6$; thứ ba là $\frac{1210}{5} = 242,0$; thứ tư là $\frac{1086}{5} = 217,2$ và thứ năm là $\frac{1440}{5} = 288,0$.

Chúng ta có thể tính Q_c theo công thức định nghĩa như sau:

$$Q_c = 5(251,0 - 252,16)^2 + 5(262,6 - 252,16)^2 + \dots + 5(288,0 - 252,16)^2 = 13601,36$$

$$Q_t = \frac{1349^2}{5} + \frac{1314^2}{5} + \dots + \frac{1188^2}{5} - \frac{6304^2}{25} = 4156,56$$

$$Q_e = 36571,36 - (6146,16 + 13601,36 + 4156,56) = 12667,28$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Khối tuần (hàng)	4	6146,16	$S^2 = 1055,611$
Khối thứ (cột)	4	13601,36	
Nghiệm thức	4	4156,56	
Sai lệch ngẫu nhiên	12	12667,28	
Tổng	24	36571,36	
$F(4,12) = \frac{\frac{4156,56}{4}}{\frac{12667,28}{12}} = 0,98; P > 0,5$			

Vậy giả thuyết tương đồng $H_0: m_1 = m_2 = \dots = m_5$ được chấp nhận, sự thăng giáng của các nghiệm thức chỉ là sự thăng giáng ngẫu nhiên, nghĩa là các nghiệm thức là như nhau: sự thích đi coi nhạc kịch là giống nhau ở mọi lứa tuổi nghiên cứu. Trong trường hợp này thì không cần phải tiến hành so sánh đối chiếu giữa các nghiệm thức với nhau nữa.

Đến đây, trước khi chuyển sang nghiên cứu mục khác, thiết nghĩ chúng ta cũng nên tổng quát hóa vấn đề là phân tích nguồn biến lượng theo kiểu “hoàn cảnh K phương” như sau: Giả sử Q_h ($h = 1, 2, k$) là tổng các bình phương cho hoàn cảnh phương h thì biểu thức phân tích nguồn biến lượng sẽ như sau:

- Đối với kiểu thu thập số liệu theo “hoàn cảnh vô phương” (đồng nhất), tức kiểu hoàn toàn ngẫu nhiên hay kiểu số liệu không tạo khối:

$$Q = Q_t + Q_e$$

- Đối với kiểu thu thập số liệu theo “hoàn cảnh nhất phương” tức kiểu số liệu kết thành khối:

$$Q = Q_1 + Q_t + Q_e, \text{ với } Q_1 = Q_b$$

- Đối với kiểu thu thập số liệu theo “hoàn cảnh nhị phương” tức kiểu bình phương Latin hay kiểu số liệu kết thành khối hàng lẫn cột:

$$Q = Q_1 + Q_2 + Q_t + Q_e, \text{ với } Q_1 = Q_r, Q_2 = Q_c$$

- Đối với kiểu thu thập số liệu theo “hoàn cảnh K phương”

$$Q = \sum_{h=1}^K Q_h + Q_t + Q_e$$

3.1.3.4. Trường hợp số liệu ở các “khối” bị thiếu

Trong nhiều trường hợp chúng ta không có được một vài số liệu ở các “khối” do không lấy được số liệu hoặc do các “khối” bố trí thí nghiệm bị hư hại hoặc bị phá hoại bởi nguyên nhân nào đó thì tiến trình phân tích nguồn biến lượng bao gồm các bước sau đây:

- *Bước 1:* Ước định số liệu bị thiếu như sau:

(i) Với một số liệu bị thiếu:

• Kiểu khối đầy đủ: Dùng công thức sau đây:

$$x = \frac{lT + rB - S}{(l-1)(r-1)}$$

Ở đây: T, B và S là tổng của nghiệm thức bị thiếu, tổng của khối bị thiếu và tổng chung, còn l và r là số nghiệm thức và số “khối” (lần lặp lại)

• Kiểu bình phương Latin : Dùng công thức sau đây:

$$x = \frac{r(R + C + T) - 2S}{(r-1)(r-2)}$$

Với nhiều (≥ 2) số liệu bị thiếu: Dùng phương pháp “cân đối qua lại” để có được ước định tối ưu cho các số liệu bị thiếu.

Bước 2: Tính Q_c trong bối cảnh của các số liệu ước định.

Bước 3: Tính Q , Q_b (hay Q_r và Q_c) với các số liệu ban đầu.

Bước 4: Tính Q_i như sau:

$$Q_i = Q - (Q_b + Q_c) \text{ (kiểu khối đầy đủ)}$$

$$Q_i = Q - (Q_r + Q_c + Q_c) \text{ (kiểu bình phương Latin)}$$

Chú ý rằng, số độ tự do, số số liệu của các nghiệm thức và trung bình của các nghiệm thức được tính từ số liệu ban đầu.

* Sau đây là ví dụ minh họa đối với kiểu khối đầy đủ bị thiếu một số liệu:

Nghiệm thức	Khối					Tổng	Trung bình
	1	2	3	4	5		
A	32,3	34,0	34,3	35,0	36,5	172,1	34,4
B	33,3	33,0	36,3	36,8	34,5	173,9	34,8
C	30,8	34,3	35,3	32,3	35,8	168,5	33,7
D	—	26,0	29,8	28,0	28,8	112,6	28,2
Tổng	96,4	127,3	135,7	132,1	135,6	627,1	33,0

- **Bước 1:** Ước định số liệu bị thiếu:

$$x = \frac{4112,6 + 5.96,4 - 627,1}{(4-1)(5-1)} = 25,4$$

- **Bước 2:** Tính Q_c trong bối cảnh của số liệu ước định:

$$Q' = 32,3^2 + 34,0^2 + \dots + 25,4^2 + \dots + 28,8^2 - \frac{(627,1 + 25,4)^2}{19 + 1} = 224,08$$

$$Q_b = \frac{(96,4 + 25,4)^2}{3 + 1} + \frac{127,3^2}{4} + \dots + \frac{135,6^2}{4} - \frac{(627,1 + 25,4)^2}{19 + 1} = 35,39$$

$$Q_t = \frac{172,1^2}{5} + \frac{173,9^2}{5} + \frac{168,5^2}{5} + \frac{(112,6 + 25,5)^2}{4+1} - \frac{(627,1 + 25,4)^2}{19+1} = 171,36$$

$$Q_e = 224,08 - (35,39 + 171,36) = 17,33$$

- Bước 3: tính Q , Q_b với các số liệu ban đầu:

$$Q = 32,3^2 + 34,0^2 + \dots + 28,8^2 - \frac{627,1^2}{19} = 169,13$$

$$Q_b = \frac{96,4^2}{3} + \frac{127,3^2}{4} + \frac{135,6^2}{4} - \frac{627,1^2}{19} = 14,44$$

- Bước 4 : tính Q_t

$$Q_t = 169,13 - (14,44 + 17,33) = 137,36$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Khối	4	(14,44)	$S^2 = 1,58$
Nghiem thức	3	137,36	
Sai lệch ngẫu nhiên	11	17,33	
Tổng	18	(169,13)	
$F(3, 11) = \frac{\frac{137,36}{3}}{\frac{17,33}{11}} = 29,06; P < 0,005$			

* Sau đây là ví dụ minh họa đối với kiểu bình phương Latin bị thiếu một số liệu:

Hàng	Cột			Tổng
	1	2	3	
I	A(-)	B(885)	C(940)	1825
II	B(715)	C(1087)	A(766)	2568
III	C(844)	A(711)	B(832)	2387
Tổng	1559	2683	2538	6780

	Thử nghiệm			
	A	B	C	
	-	715	844	Tổng
	711	885	1087	
	766	832	940	
Tổng	1477	2432	2871	6780
Trung bình	738,5	810,7	957	847,5

Bước 1: Ước định số liệu bị thiếu:

$$x = \frac{3(1825 + 1559 + 1477) - 2 \cdot 6780}{(3-1)(3-2)} = 512$$

Bước 2: Tính Q_c trong bối cảnh của số liệu ước định:

$$Q' = 512^2 + 885^2 + \dots + 832^2 + \frac{(6780 + 512)^2}{8+1} = 210459,6$$

$$Q_r = \frac{(1825 + 512)^2}{2+1} + \frac{2568^2}{3} + \frac{2387^2}{3} - \frac{(6780 + 512)^2}{8+1} = 9846,9$$

$$Q_c = \frac{(1559 + 512)^2}{2+1} + \frac{2683^2}{3} + \frac{2538^2}{3} - \frac{(6780 + 512)^2}{8+1} = 68184,2$$

$$Q_t = \frac{(1477 + 512)^2}{2+1} + \frac{2432^2}{3} + \frac{2871^2}{3} - \frac{(6780 + 512)^2}{8+1} = 129654,9$$

$$Q_e = 210459,6 - (9846,9 + 68184,2 + 129654,9) = 2773,6$$

Bước 3: Tính Q , Q_r , Q_c với các số liệu ban đầu:

$$Q = 885^2 + 940^2 + \dots + 832^2 - \frac{6780^2}{8} = 110406$$

$$Q_r = \frac{1825^2}{2} + \frac{2568^2}{3} + \frac{2387^2}{3} - \frac{6780^2}{8} = 16726,8$$

$$Q_c = \frac{1559^2}{2} + \frac{2683^2}{3} + \frac{2538^2}{3} - \frac{6780^2}{8} = 15834,8$$

Bước 4: Tính Q_t :

$$Q_t = 110406 - (16726,8 + 15834,8 + 2773,6) = 75070,8$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Khối hàng	2	(16726,8)	$S^2 = 2773,6$
Khối cột	2	(15834,8)	
Nghiem thức	2	75070,8	
Sai lệch ngẫu nhiên	1	2773,6	
Tổng	7	(110406)	

$$F(2,1) = \frac{\frac{75070,8}{2}}{\frac{2773,6}{1}} = 13,5; \quad P > 0,1$$

* Sau đây là ví dụ minh họa đối với kiểu khối đầy đủ bị thiếu hai số liệu:

Nghiem thức	Khối						Tổng	Trung bình
	1		2		3			
A	6		5		4		15	5
B	15		-		8		23	11,5
C	-		15		12		27	13,5
Tổng	21		20		24		65	9,3

Bước 1: Ước định các số liệu bị thiếu:

Trước hết cần ước định hợp lý sơ bộ cho một số liệu bị thiếu, chẳng hạn, với x_{22} chúng ta thấy rằng trung bình của khối là 10 còn trung bình nghiệm thức là 11,5, do đó ước định hợp lý sơ bộ cho x_{22} có thể là $x_{22} = 10,5$. Bây giờ, nếu $x_{22} = 10,5$ thì x_{31} sẽ là:

$$x_{31} = \frac{3(27) + 3(21) - (65 + 10,5)}{(3-1)(3-1)} = 17,1$$

+ nếu $x_{31} = 17,1$ thì x_{22} sẽ là:

$$x_{22} = \frac{3(23) + 3(20) - (65 + 17,1)}{(3-1)(3-1)} = 11,7$$

+ nếu $x_{22} = 11,7$ thì x_{31} sẽ là:

$$x_{31} = \frac{3(27) + 3(21) - (65 + 11,7)}{(3-1)(3-1)} = 16,8$$

+ nếu $x_{31} = 16,8$ thì x_{22} sẽ là:

$$x_{22} = \frac{3(23) + 3(20) - (65 + 16,8)}{(3-1)(3-1)} = 11,8$$

+ nếu $x_{22} = 11,8$ thì x_{31} sẽ là:

$$x_{31} = \frac{3(27) + 3(21) - (65 + 11,8)}{(3-1)(3-1)} = 16,8 \text{ (không đổi)}$$

Vậy các số liệu bị thiếu được ước định là $x_{22} = 11,8$ và $x_{31} = 16,8$

Bước 2: Tính Q_e trong bối cảnh của các số liệu ước định:

$$Q' = 6^2 + \dots + 11,8^2 + 16,8^2 + \dots + 12^2 - \frac{(65 + 11,8 + 16,8)^2}{7+2} = 183,04$$

$$Q_b = \frac{(21+16,8)^2}{2+1} + \frac{(20+11,8)^2}{2+1} + \frac{24^2}{3} - \frac{(65+11,8+16,8)^2}{7+2} = 31,92$$

$$Q_t = \frac{15^2}{3} + \frac{(23+11,8)^2}{2+1} + \frac{(27+16,8)^2}{2+1} - \frac{(65+11,8+16,8)^2}{7+2} = 144,72$$

$$Q_e = 183,04 - (31,92 + 144,72) = 6,4$$

Bước 3: Tính Q , Q_b với các số liệu ban đầu:

$$Q = 6^2 + 5^2 + \dots + 12^2 - \frac{65^2}{7} = 131,43$$

$$Q_b = \frac{21^2}{2} + \frac{20^2}{2} + \frac{24^2}{3} - \frac{65^2}{7} = 8,93$$

Bước 4: Tính Q_t

$$Q_t = 131,43 - (8,93 + 6,4) = 116,1$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Khối	2	(8,93)	S ² = 3,2
Nghiệm thức	2	116,1	
Sai lệch ngẫu nhiên	2	6,4	
Tổng	6	(131,43)	

$$F(2,2) = \frac{\frac{116,1}{2}}{\frac{6,4}{2}} = 18,14; P > 0,05$$

* Sau đây là ví dụ minh họa đối với kiểu bình phương Latin bị thiếu hai số liệu:

Hàng	Cột				Tổng
	1	2	3	4	
I	A(-)	B(8)	C(17)	D(15)	40
II	B(6)	C(14)	D(16)	A(9)	46
III	C(10)	D(-)	A(7)	B(12)	29
	D(15)	A(9)	B(16)	C(19)	59
Tổng	25	42		2538	6780

	Thử nghiệm				
	A	B	C	D	
	-	6	10	15	
	9	8	14	-	
	7	16	17	16	
	9	12	19	15	
Tổng	25	42	60	46	173

Bước 1: Ước định các số liệu bị thiếu: Ước định hợp lý sơ bộ và cân đối qua lại như trên nhưng dùng công thức của trường hợp kiểu bình phương Latin, cuối cùng chúng ta có được kết quả là: $x_{111} = 2,2$ và $x_{234} = 12,3$.

Bước 2: Tính Q_e trong bối cảnh của các số liệu ước định:

$$Q' = 2,2^2 + 8^2 + \dots + 12,3^2 + \dots + 19^2 - \frac{(173 + 2,2 + 12,3)^2}{14 + 2} = 321,86$$

$$Q_c = \frac{(31 + 2,2)^2}{3 + 1} + \frac{(31 + 12,3)^2}{3 + 1} + \frac{56^2}{4} + \frac{55^2}{4} - \frac{(173 + 2,2 + 12,3)^2}{14 + 2} = 87,27$$

$$Q_r = \frac{(40 + 2,2)^2}{3 + 1} + \frac{45^2}{4} + \frac{(29 + 12,3)^2}{3 + 1} + \frac{59^2}{4} - \frac{(173 + 2,2 + 12,3)^2}{14 + 2} = 50,87$$

$$Q_t = \frac{(25 + 2,2)^2}{3 + 1} + \frac{42^2}{4} + \frac{60^2}{4} + \frac{(46 + 12,3)^2}{3 + 1} - \frac{(173 + 2,2 + 12,3)^2}{14 + 2} = 178,42$$

$$Q_e = 321,86 - (87,27 + 50,87 + 178,42) = 5,3$$

Bước 3: Tính Q , Q_r , Q_c với các số liệu ban đầu:

$$Q = 8^2 + 17^2 + \dots + 19^2 - \frac{173^2}{14} = 25,21$$

$$Q_r = \frac{40^2}{3} + \frac{45^2}{4} + \frac{29^2}{3} + \frac{59^2}{4} - \frac{173^2}{14} = 52,38$$

$$Q_c = \frac{31^2}{3} + \frac{31^2}{3} + \frac{56^2}{4} + \frac{55^2}{4} - \frac{173^2}{14} = 43,13$$

Bước 4: Tính Q_t :

$$Q_t = 225,21 - (52,38 + 43,13 + 5,3) = 124,4$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Khối hàng	3	(52,38)	$S^2 = 1,33$
Khối cột	3	(43,13)	
Nghiệm thức	3	124,40	
Sai lệch ngẫu nhiên	4	5,30	
Tổng	13	(225,21)	
$F(3,4) = \frac{\frac{124,4}{3}}{\frac{5,3}{4}} = 31,3; P < 0,005$			

Sự phân tích nguồn biến động cho trường hợp có nhiều (> 2) số liệu bị thiếu cũng tiến hành tương tự như trường hợp có hai số liệu bị thiếu. Tất nhiên, càng có nhiều số liệu bị thiếu thì kết quả so sánh các nghiệm thức càng kém chính xác, do đó nếu có thể được tốt hơn hết là thu thập lại số liệu hoặc bố trí lại các thí nghiệm nghiên cứu.

3.2. SO SÁNH CÁC NGHIỆM THỨC DẠNG TỶ LỆ

3.2.1. So sánh một tỷ lệ với một tỷ lệ có sẵn

Cơ sở khoa học của thuật so sánh đã được trình bày đầy đủ ở chương 2, do đó ở đây sẽ đưa ra các ví dụ minh họa cùng các phương pháp khác nhau được áp dụng để thực hiện việc so sánh các nghiệm thức với nhau.

Ví dụ 3.1. Tỷ lệ phế phẩm cho phép theo quy định chung là 12%. Kiểm tra 200 sản phẩm của nhà máy thấy có 30 phế phẩm. Hỏi tỷ lệ phế phẩm của

nhà máy này có vượt quá tỷ lệ phế phẩm cho phép theo quy định chung hay không ?

Cách 1: Đặt giả thuyết tương đồng $H_0: p = 0,12$. Vì rằng $np = 200 \cdot 0,12 = 24 > 5$, do đó:

$$U = \frac{\frac{30}{200} - 0,12}{\sqrt{\frac{0,12(1-0,12)}{200}}} = 1,31; \quad P > 0,1$$

Vậy giả thuyết tương đồng $H_0: p = 0,12$ được chấp nhận, sự thăng giáng của tỷ lệ $\frac{30}{200}$ chỉ là sự thăng giáng ngẫu nhiên của tỷ lệ $\frac{12}{100}$, nghĩa là tỷ lệ phế phẩm của nhà máy không vượt quá tỷ lệ phế phẩm cho phép theo quy định chung.

Cách 2: Với $P = 0,05$ thì $U_p = U_{0,05} = 1,96$, do đó:

$$\begin{aligned} -U_p < U < U_p &\Rightarrow -1,96 < \frac{\frac{30}{200} - p}{\sqrt{\frac{\frac{30}{200} \left(1 - \frac{30}{200}\right)}{200}}} < 1,96 \Rightarrow \\ &\Rightarrow \frac{30}{200} - 1,96 \cdot \sqrt{\frac{\frac{30}{200} \left(1 - \frac{30}{200}\right)}{200}} < p < \frac{30}{200} + 1,96 \cdot \sqrt{\frac{\frac{30}{200} \left(1 - \frac{30}{200}\right)}{200}} \\ &\Rightarrow 0,15 - 0,0495 < p < 0,15 + 0,0495 \Rightarrow 0,1005 < p < 0,1995 \end{aligned}$$

Vì $0,12 \in (0,1005; 0,1995)$, do đó, tỷ lệ phế phẩm của nhà máy và tỷ lệ phế phẩm quy định là như nhau:

- *Cách 3:*

Tri thực tế

30

170

n = 200

Tri lý thuyết

$200 \cdot \frac{12}{100} = 24$

$200 \cdot \frac{88}{100} = 176$

n = 200

$$\chi^2(2-1) = \frac{(30-24)^2}{24} + \frac{(170-176)^2}{176} = 1,71; P > 0,1$$

(Kết luận như trên).

Trong trường hợp này chúng ta có thể tính nhanh giá trị χ^2 như sau:

$$\chi^2(1) = \frac{\left[30 - 170 \cdot \left(\frac{12}{88} \right) \right]^2}{200 \cdot \frac{12}{88}} = 1,71$$

hay

$$\chi^2(1) = \frac{\left[170 - 30 \cdot \left(\frac{88}{12} \right) \right]^2}{200 \cdot \frac{88}{12}} = 1,71$$

Để nhận thấy: $U^2 = 1,31^2 = 1,71 = \chi^2(1)$

* *Chú ý:* Nếu gặp trường hợp $np < 5$ thì xử lý như sau:

(i) Dùng trắc nghiệm χ^2 có tu chỉnh (xem chương 2).

(ii) Nếu p quá nhỏ, dẫn tới np quá nhỏ, thì cần lấy thật nhiều mẫu rồi áp dụng công thức xác định giới hạn tin tưởng dựa căn bản trên phân phối Poisson đã trình bày ở chương 2 để tiến hành việc so sánh (giống như cách 2 ở trên).

3.2.2. So sánh hai tỷ lệ với nhau

Ví dụ 3.2: So sánh kết quả điều trị bệnh theo hai phương pháp khác nhau sau đây:

Kết quả	Phương pháp điều trị		Tổng
	A	B	
Khỏi bệnh	85	175	260
Không khỏi bệnh	15	25	40
Tổng	100	200	300

Giả thuyết tương đồng $H_0: p_1 = p_2$

$$\min [n_i p_i] = \frac{100 \cdot 40}{300} = 13,3 > 5$$

Cách 1:

$$U = \frac{\left(\frac{85}{100} - \frac{175}{200}\right) - 0}{\sqrt{\frac{260}{300} \cdot \frac{40}{300} \left(\frac{1}{100} + \frac{1}{200}\right)}} = -0,6; P > 0,5$$

Vậy giả thuyết $H_0: p_1 = p_2$ được chấp nhận, sự thăng giáng của các tỷ lệ $\frac{85}{100}$ và $\frac{175}{200}$ chỉ là sự thăng giáng ngẫu nhiên, nghĩa là kết quả điều trị bệnh của hai phương pháp là như nhau.

Cách 2: Với $P = 0,05$ thì $U_p = U_{0,05} = 1,96$, do đó:

$$\begin{aligned} -1,96 &< \frac{\left(\frac{85}{100} - \frac{175}{200}\right) - (p_1 - p_2)}{\sqrt{\frac{260}{300} \cdot \frac{40}{300} \left(\frac{1}{100} + \frac{1}{200}\right)}} < 1,96 \Rightarrow \\ \Rightarrow \left(\frac{85}{100} - \frac{175}{200}\right) - 1,96 \sqrt{\frac{260}{300} \cdot \frac{40}{300} \left(\frac{1}{100} + \frac{1}{200}\right)} &< p_1 - p_2 < \\ < \left(\frac{85}{100} - \frac{175}{200}\right) + 1,96 \sqrt{\frac{260}{300} \cdot \frac{40}{300} \left(\frac{1}{100} + \frac{1}{200}\right)} \\ \Rightarrow 0,025 - 0,0816 < p_1 - p_2 < -0,025 + 0,0816 \\ \Rightarrow -0,1066 < p_1 - p_2 < 0,0566. \end{aligned}$$

Vì $(p_1 - p_2) \in (-0,1066; 0,0566)$, khoảng tin tưởng này bao gồm giá trị zero, do đó $p_1 - p_2 \sim 0 \Rightarrow p_1 \sim p_2$ (kết luận như trên).

Cách 3:

<u>Tri thức tế</u>	<u>Tri lý thuyết</u>
85	$\frac{100 \cdot 260}{300} = 86,67$
15	$\frac{100 \cdot 40}{300} = 13,33$
175	$\frac{200 \cdot 260}{300} = 173,33$
25	$\frac{200 \cdot 40}{300} = 26,67$
<u>n = 300</u>	<u>n = 300</u>

$$\chi^2[(2-1)(2-1)] = \frac{(85-86,67)^2}{86,67} + \frac{(15-13,33)^2}{13,33} + \frac{(175-173,33)^2}{173,33} + \frac{(25-26,67)^2}{26,67} = 0,36; P > 0,5$$

(Kết luận như trên).

Để nhận thấy là $U^2 = (-0,6)^2 = 0,36 = \chi^2(1)$. Thực ra, chúng ta có thể tính nhanh giá trị χ^2 như sau:

$$\chi^2[(2-1)(2-1)] = 300 \left(\frac{85^2}{100.260} + \frac{15^2}{100.40} + \frac{175^2}{200.260} + \frac{25^2}{200.40} - 1 \right) = 0,36$$

Ngoài ra, đối với bảng 2×2 thường tính nhanh giá trị χ^2 , như sau:

$$\chi^2[(2-1)(2-1)] = \frac{300(85.25 - 15.175)^2}{100.200.260.40} = 0,36$$

3.2.3. So sánh nhiều tỷ lệ với nhau

Dùng trắc nghiệm χ^2 để so sánh chung cho nhiều tỷ lệ với nhau. Một khi giả thuyết tương đồng $H_0: p_1 = p_2 = \dots = p_k$ bị bác bỏ ($P \leq 0,05$) thì có nghĩa là trong K tỷ lệ nghiệm thức có ít nhất là 2 tỷ lệ thực sự khác nhau, khi đó để xác định các tỷ lệ nào là khác nhau chúng ta thực hiện so sánh đối chiếu cho từng cặp nghiệm thức bằng trắc nghiệm χ^2 với 1 độ tự do, bằng trắc nghiệm U hay bằng cách dựa vào khoảng tin tưởng giống như khi so sánh hai tỷ lệ nghiệm thức với nhau.

Ví dụ 3.3: So sánh kết quả điều trị bệnh theo 3 phương pháp khác nhau sau đây:

Kết quả	Phương pháp điều trị			Tổng
	A	B	C	
Khỏi bệnh	105	86	120	311
Không khỏi bệnh	55	34	30	119
Tổng	160	120	150	430
% Khỏi bệnh	65,6	71,7	80,0	

Giả thuyết tương đồng $H_0: p_1 = p_2 = p_3$

$$\min [n_i p_i] = 120 \cdot \frac{119}{430} = 33,2 > 5$$

$$\chi^2[(2-1)(3-1)] = 430 \left(\frac{105^2}{160 \cdot 311} + \frac{86^2}{120 \cdot 311} + \dots + \frac{30^2}{150 \cdot 119} - 1 \right) = 8,03;$$

$$P < 0,025.$$

Vậy giả thuyết H_0 bị bác bỏ, nghĩa là có ít nhất là 2 trong số các tỷ lệ nghiệm thức này là thăng giáng thực sự. Do đó, chúng ta cần phải tiến hành các so sánh đối chiếu sau đây:

So sánh đối chiếu A/B:

A	B	Tổng	giả thuyết $H_0: p_1 = p_2$
105	86	191	$\min [n_i p_i] = \frac{120 \cdot 89}{280} = 38,1 > 5$
55	34	89	
Tổng	160	120	

$$\chi^2[(2-1)(3-1)] = \frac{280(105 \cdot 34 - 55 \cdot 86)^2}{160 \cdot 120 \cdot 191 \cdot 89} = 1,15; P > 0,25$$

Vậy giả thuyết $H_0: p_1 = p_2$ được chấp nhận, sự thăng giáng của các tỷ lệ nghiệm thức A và B chỉ là sự thăng giáng ngẫu nhiên, nghĩa là kết quả điều trị của các phương pháp A và B là như nhau.

So sánh đối chiếu A/C :

A	C	Tổng	giả thuyết $H_0: p_1 = p_3$
105	120	225	$\min [n_i p_i] = \frac{150 \cdot 85}{310} = 41,1 > 5$
55	30	85	
Tổng	160	150	310

$$\chi^2[(2-1)(2-1)] = \frac{310(105 \cdot 30 - 55 \cdot 120)^2}{160 \cdot 150 \cdot 225 \cdot 89} = 8,04; P < 0,005$$

Vậy giả thuyết $H_0: p_1 = p_3$ bị bác bỏ, sự thăng giáng của các tỷ lệ nghiệm thức A và C là một sự thăng giáng thực sự, nghĩa là kết quả điều trị của các phương pháp A và C là thực sự khác nhau.

So sánh đối chiếu B/C:

B	C	Tổng	giả thuyết $H_0: p_2 = p_3$
86	120	206	

	34	30	64	
Tổng	120	150	270	$\min [n_i p_i] = \frac{120 \cdot 64}{270} = 28,4 > 5$

$$\chi^2 [(2-1)(2-1)] = \frac{270(86 \cdot 30 - 34 \cdot 120)^2}{120 \cdot 150 \cdot 206 \cdot 64} = 2,56; P > 0,1$$

Vậy giả thuyết $H_0: p_2 = p_3$ được chấp nhận, sự thăng giáng của các tỷ lệ nghiệm thức B và C chỉ là sự thăng giáng ngẫu nhiên, nghĩa là kết quả điều trị của các phương pháp B và C là như nhau.

Chú ý rằng, đối với các đối chiếu A/B, A/C và B/C thay vì dùng trắc nghiệm χ^2 chúng ta có thể dùng khoảng tin tưởng hay trắc nghiệm U để thực hiện việc so sánh giống như việc so sánh hai tỷ lệ đã được trình bày ở trên. Chẳng hạn, đối với so sánh đối chiếu A/C thì:

$$U = \frac{\left(\frac{105}{160} - \frac{120}{150}\right) - 0}{\sqrt{\frac{225}{310} \cdot \frac{85}{310} \left(\frac{1}{160} + \frac{1}{150}\right)}} = -2,835; P < 0,005 \text{ (kết luận như trên).}$$

Để nhận thấy: $U^2 = (-2,835)^2 = 8,04 = \chi^2(1)$.

Hoặc dựa vào khoảng tin tưởng như sau:

$$\left(\frac{105}{160} - \frac{120}{150}\right) - 1,96 \sqrt{\frac{225}{310} \cdot \frac{85}{310} \left(\frac{1}{160} + \frac{1}{150}\right)} < p_1 - p_3 < \left(\frac{105}{160} - \frac{120}{150}\right) + 1,96 \sqrt{\frac{225}{310} \cdot \frac{85}{310} \left(\frac{1}{160} + \frac{1}{150}\right)}$$

$$\Rightarrow -0,14375 - 0,00257 < p_1 - p_3 < -0,14375 + 0,00257$$

$\Rightarrow -0,14632 < p_1 - p_3 < -0,14118$. Vì khoảng tin tưởng của $(p_1 - p_3)$ là $(-0,14632; -0,14118)$ không bao gồm giá trị zezo, do đó, $p_1 - p_3 \neq 0 \Rightarrow p_1 \neq p_3$ (kết luận như trên).

3.3. SO SÁNH HAI NGHIỆM THỨC VỚI NHAU

3.3.1. Với số liệu không bất “cấp”

3.3.1.1. Thuật so sánh phi tham số (non - parametric test)

Trong trường hợp số liệu không bất “cấp” hay không tạo “khối” (kiểu hoàn toàn ngẫu nhiên) thì các phương pháp so sánh phi tham số (non-parametric methods) thường dùng là các trắc nghiệm về hạng (rank tests) mà phổ biến nhất

là trắc nghiệm Wilcoxon - Mann - whitney, thường được mệnh danh là trắc nghiệm Wilcoxon (Wilcoxon's test). Nội dung của trắc nghiệm này như sau: Giả sử nghiệm thức I có n_1 số liệu, nghiệm thức II có n_2 số liệu, sắp xếp $(n_1 + n_2)$ số liệu của 2 nghiệm thức từ nhỏ tới lớn. Đối với từng số liệu của mỗi nghiệm thức chúng ta xem trước nó có bao nhiêu số liệu của nghiệm thức kia và tính tổng w_1 và w_2 của chúng. Vì rằng $w_1 + w_2 = n_1 n_2$ nên công thức này được dùng để kiểm tra việc tính toán. Về phương diện thực hành các giá trị w_1 và w_2 thường được tính theo các công thức sau đây:

$$w_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

và
$$w_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

Ở đây, R_1 và R_2 là tổng hạng của nghiệm thức I và nghiệm thức II.

Để minh họa cho việc tính toán chúng ta xét một trường hợp cụ thể sau đây:

Nghiệm thức I	39; 50; 61; 67; 40; 40; 54
Nghiệm thức II	60; 53; 42; 41; 40; 54; 63; 69
Số liệu	$\overline{39}; \overline{40}; \overline{40}; 40; 41; 42; \overline{50}; 53; \overline{54}; 54; 60; \overline{61}; 63; \overline{67}; 69$
Hạng	1; 3; 3; 3; 5; 6; 7; 8; 9,5; 9,5; 11; 12; 13; 14; 15

$$R_1 = 1 + 3 + 3 + 7 + 9,5 + 12 + 14 = 49,5; n_1 = 7$$

$$R_2 = 3 + 5 + 6 + 8 + 9,5 + 11 + 13 + 15 = 70,5; n_2 = 8$$

$$w_1 = 7.8 + \frac{7(7+1)}{2} - 49,5 = 34,5$$

$$w_2 = 7.8 + \frac{8(8+1)}{2} - 70,5 = 21,5$$

Rõ ràng là: $34,5 + 21,5 = 7.8 = 56$

$$(w_1) \quad (w_2) \quad (n_1) \quad (n_2)$$

Thực ra, chúng ta có thể tính w_1 và w_2 theo định nghĩa như sau:

$$w_1 = 2 + 3 + 3 + 4 + 4,5 + 5 + 6 + 7 = 34,5$$

$$w_2 = 0 + 0,5 + 0,5 + 3 + 4,5 + 6 + 7 = 21,5$$

Chú ý rằng, nếu sự “thăng giáng nghiệm thức” là sự “thăng giáng ngẫu nhiên” thì $w_1 \sim w_2$, còn nếu sự “thăng giáng nghiệm thức” là một sự thăng giáng thực sự thì $w_1 \neq w_2$.

Vì rằng đại lượng $U = \frac{\omega - \frac{1}{2}n_1n_2}{\sqrt{\frac{1}{12}n_1n_2(n_1+n_2+1)}}$ có phân phối cận bình

thường dạng chuẩn $N(0, 1)$, nên chúng ta có thể dùng trắc nghiệm U để kiểm tra giả thuyết tương đồng $H_0: w_1 = w_2$. Chú ý rằng, thực ra nếu như w lấy trị w_1 thì giả thuyết tương đồng phải là $H_0: w_1 = \frac{1}{2}n_1n_2$, còn nếu như w lấy trị w_2 thì giả thuyết tương đồng phải là $H_0: w_1 = \frac{1}{2}n_1n_2$. Tuy vậy, ba cách đặt giả thuyết H_0 này là như nhau, bởi vì như chúng ta đã biết là: $w_1 + w_2 = n_1n_2$, do đó, nếu như giả thuyết $H_0: w_1 = \frac{1}{2}n_1n_2$ được chấp nhận thì $w_2 = n_1n_2 - w_1 = n_1n_2 - \frac{1}{2}n_1n_2 = \frac{1}{2}n_1n_2 \Rightarrow w_1 = w_2$, còn nếu như giả thuyết $H_0: w_2 = \frac{1}{2}n_1n_2$ được chấp nhận thì $w_1 = n_1n_2 - w_2 = n_1n_2 - \frac{1}{2}n_1n_2 = \frac{1}{2}n_1n_2 \Rightarrow w_1 = w_2$.

Ví dụ 3.4: Đánh giá kết quả tăng trưởng về chiều cao (cm) của cây con được gieo ươm trên hai loại đất khác nhau (A và B) (xem lại ví dụ 2.(19)):

A	50; 41; 48; 60; 46; 60; 51; 42; 62; 54; 42; 46
B	38; 40; 47; 51; 63; 50; 63; 57; 59; 51

Số liệu	38; 40; $\overline{41}$; $\overline{42}$; $\overline{42}$; $\overline{46}$; $\overline{46}$; 47; $\overline{48}$; $\overline{50}$; 50; $\overline{51}$; 51; 51; $\overline{54}$;
Hạng	1; 2; 3; 4,5; 4,5; 6,5; 6,5; 8; 9; 10,5; 10,5; 13; 13; 13; 15;

Số liệu	57; 59; $\overline{60}$; $\overline{60}$; $\overline{62}$; 63; 63
Hạng	16 27 18,5 18,5 20 21,5 21,5

$$R_1 = 3 + 4,5 + \dots + 20 = 129,5; n_1 = 12$$

$$R_2 = 1 + 2 + \dots + 21,5 = 123,5; n_2 = 10$$

$$w_1 = 12 \cdot 10 + \frac{12(12+1)}{2} - 129,5 = 68,5$$

$$w_2 = 12 \cdot 10 + \frac{10(10+1)}{2} - 123,5 = 51,5$$

Dễ nhận thấy là: $68,5 + 51,5 = 12 \cdot 10 = 120$

Giả thuyết tương đồng $H_0: w_1 = w_2$ (hay $H_0: w_1 = \frac{1}{2} n_1 n_2$)

$$\text{Vì rằng: } U = \frac{51,5 - \frac{1}{2} \cdot 12 \cdot 10}{\sqrt{\frac{1}{12} \cdot 12 \cdot 10(12+10+1)}} = 0,56; P > 0,5 \text{ (kết luận như trên).}$$

Vậy giả thuyết H_0 được chấp nhận, sự thăng giáng nghiệm thức chỉ là sự thăng giáng ngẫu nhiên, nghĩa là các nghiệm thức A và B là như nhau.

Chú ý rằng, nếu như chúng ta đặt giả thuyết tương đồng là:

$$H_0: w_1 = \frac{1}{2} n_1 n_2 \text{ thì:}$$

$$U = \frac{68,5 - \frac{1}{2} \cdot 12 \cdot 10}{\sqrt{\frac{1}{12} \cdot 12 \cdot 10(12+10+1)}} = 0,56; P > 0,5 \text{ (kết quả như trên).}$$

3.3.1.2. Các phương pháp so sánh khác

3.3.1.2.(a). Dùng trắc nghiệm t

Đây là trắc nghiệm thông dụng nhất, thường được dùng cho trường hợp so sánh hai nghiệm thức với nhau. Để dễ hình dung vấn đề, chúng ta vẫn lấy lại số liệu ở ví dụ 3.4. Vấn đề đã được trình bày ở ví dụ 2.(19), do đó, ở đây chỉ xin nhắc lại như sau:

$$A: \bar{x}_1 = 50,17; S_1^2 = 55,06; n_1 = 12$$

$$B: \bar{x}_2 = 51,90; S_2^2 = 76,32; n_2 = 10$$

Trước hết cần tiến hành trắc nghiệm so sánh các biến lượng như sau:

$$\text{Cách 1: } H_0: \sigma_1^2 = \sigma_2^2 (= \sigma^2)$$

$$F(n_2 - 1, n_1 - 1) = \frac{\frac{S_2^2}{\sigma_2^2}}{\frac{S_1^2}{\sigma_1^2}} = \frac{S_2^2}{S_1^2} \Rightarrow F(9, 11) = \frac{76,32}{55,06} = 1,39; P > 0,1$$

Vậy giả thuyết $H_0: \sigma_1^2 = \sigma_2^2 (= \sigma^2)$ được chấp nhận, sự thăng giáng của

biến lượng S_1^2 và S_2^2 chỉ là sự thăng giáng ngẫu nhiên nên ước lượng tốt nhất của σ^2 là S^2 như sau:

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Cách 2:

$$F_{1-P}(n_1 - 1, n_2 - 1) < F(n_1 - 1, n_2 - 1) < F_P(n_1 - 1, n_2 - 1)$$

$$\Rightarrow F_{1-P}(n_1 - 1, n_2 - 1) < \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} < F_P(n_1 - 1, n_2 - 1)$$

$$\Rightarrow \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_P(n_1 - 1, n_2 - 1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{1-P}(n_1 - 1, n_2 - 1)}$$

$$\Rightarrow \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_P(n_1 - 1, n_2 - 1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} \cdot F_P(n_2 - 1, n_1 - 1)$$

Với $P = 0,05$ thì $F_P(n_1 - 1, n_2 - 1) = F_{0,05}(9, 11) = 3,11$

và $F_P(n_2 - 1, n_1 - 1) = F_{0,05}(9, 11) = 2,90$, do đó:

$$\frac{55,06}{76,32} \cdot \frac{1}{3,11} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{55,06}{76,32} \cdot 2,90 \Rightarrow 0,232 < \frac{\sigma_1^2}{\sigma_2^2} < 2,092$$

Vì khoảng tin tưởng của $\frac{\sigma_1^2}{\sigma_2^2}$ là $(0,232; 2,092)$ bao gồm giá trị 1, nghĩa là

$$\frac{\sigma_1^2}{\sigma_2^2} \sim 1 \Rightarrow \sigma_1^2 \sim \sigma_2^2 \text{ (kết quả như trên).}$$

Bây giờ chúng ta thực hiện so sánh nghiệm thức A với nghiệm thức B bằng trắc nghiệm t như sau:

Giả thuyết $H_0: m_1 = m_2$.

$$t(n_1 + n_2 - 2) = \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$\Rightarrow t(20) = \frac{(50,17 - 51,90) - 0}{\sqrt{\frac{(12-1)55,06 + (10-1)76,32}{12+10-2} \cdot \left(\frac{1}{12} + \frac{1}{10}\right)}} = -0,5; P > 0,5$$

(Kết luận giống như ở trắc nghiệm Wilcoxon).

3.3.1.2.(b). Dùng khoảng tin tưởng

Vẫn lấy số liệu ở ví dụ 3.4, chúng ta tiến hành xác định khoảng tin tưởng của $(m_1 - m_2)$ như sau:

$$\begin{aligned} & -t_p(n_1 + n_2 - 2) < t(n_1 + n_2 - 2) < t_p(n_1 + n_2 - 2) \\ \Rightarrow -t_p(n_1 + n_2 - 2) & < \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < t_p(n_1 + n_2 - 2) \\ \Rightarrow (\bar{x}_1 - \bar{x}_2) - t_p(n_1 + n_2 - 2) \cdot \sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} & < m_1 - m_2 < \\ < (\bar{x}_1 - \bar{x}_2) + t_p(n_1 + n_2 - 2) \cdot \sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \end{aligned}$$

Với $P = 0,05$ thì $t_p(n_1 + n_2 - 2) = t_{0,05}(20) = 2,086$, do đó:

$$\begin{aligned} (50,17 - 51,90) - 2,086 \cdot \sqrt{\frac{(12-1)55,06 + (10-1)76,32}{12+10-2} \cdot \left(\frac{1}{12} + \frac{1}{10}\right)} & < m_1 - \\ -m_2 < (50,17 - 51,90) + 2,086 \cdot \sqrt{\frac{(12-1)55,06 + (10-1)76,32}{12+10-2} \cdot \left(\frac{1}{12} + \frac{1}{10}\right)} \end{aligned}$$

$\Rightarrow -1,73 - 7,18 < m_1 - m_2 < -1,73 + 7,18 \Rightarrow -8,91 < m_1 - m_2 < 5,45$.
Vì khoảng tin tưởng của $(m_1 - m_2)$ là $(-8,91; 5,45)$ bao gồm giá trị zero, nghĩa là $m_1 - m_2 \sim 0 \Rightarrow m_1 \sim m_2$ (kết luận giống như ở trắc nghiệm Wilcoxon và trắc nghiệm t). Chú ý rằng, khoảng tin tưởng phụ thuộc vào trường hợp biến lượng bằng nhau hay khác nhau (giống như ở trắc nghiệm t), do đó trước khi tiến hành xác định khoảng tin tưởng của $(m_1 - m_2)$ nhất thiết phải thực hiện trắc nghiệm về các biến lượng (xem lại mục 2.3.2.).

3.3.1.2.(c). Dùng trắc nghiệm F

Vì rằng $t^2(K) = F(1, K)$, do đó, đối với số liệu ở ví dụ 3.4 chúng ta có thể

tiến hành phân tích nguồn biến động và dùng trắc nghiệm F với giả thuyết $H_0: m_1 = m_2$ như sau:

$$n = n_1 + n_2 = 12 + 10 = 22; \sum x_{i1} = 50 + 41 + \dots + 46 = 602;$$

$$\sum x_{i2} = 38 + 40 + \dots + 51 = 519; \sum x_i = 50 + 41 + \dots + 51 = 1121;$$

$$\sum x_i^2 = 50^2 + 41^2 + \dots + 51^2 = 58429$$

$$Q = 58429 - \frac{1121^2}{22} = 1308.95$$

$$Q = \frac{602^2}{12} + \frac{519^2}{10} - \frac{1121^2}{22} = 16.39$$

$$Q_e = 1308.95 - 13.39 = 1292.56; S^2 = 1292.56/20 = 64.63$$

$$F(1, 20) = \frac{\frac{16.39}{1292.56}}{20} = 0.25; P > 0.5 \text{ (kết luận như trên)}$$

Dễ nhận thấy: $t^2(20) = (-0.5)^2 = 0.25 = F(1, 20)$ và

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(12 - 1)55.06 + (10 - 1)76.32}{12 + 10 - 2} = 64.63 = \frac{Q_e}{20}$$

3.3.2. Với số liệu bất “cặp” nhau

3.3.2.1. Thuật so sánh phi tham số

Trong trường hợp số liệu bất “cặp” nhau hay kết thành “khối” (kiểu khối đầy đủ) thì các phương pháp so sánh phi tham số thường dùng là các trắc nghiệm về dấu (Sign tests) mà phổ biến nhất là trắc nghiệm Dixon - Mood (Dixon - Mood's test). Nội dung của trắc nghiệm này như sau: Để so sánh hai nghiệm thức I và II với n cặp số liệu (x_{i1}, x_{i2}) bất “cặp” nhau chúng ta lập hiệu $(x_{i1} - x_{i2})$ của từng cặp để biết kết quả là + hay - . Gọi l_1 là tổng số dấu (+ hay -) có ít hơn, và l_2 là tổng số dấu (+ hay -) có nhiều hơn và l là tổng số dấu (+ và -), tức $l = l_1 + l_2$. Chú ý rằng, nếu sự thắng giáng nghiệm thức là sự thắng giáng ngẫu nhiên thì số dấu (+) hay số dấu (-) sẽ có phân phối nhị thức dạng $B(l, \frac{1}{2})$, do đó, giả thuyết tương đồng được đặt ra là $H_0: p = \frac{1}{2}$ hoặc $H_0: l_1 = l_2$.

Vì rằng đại lượng:

$$F[2(l_1 + 1), 2l_2] = \frac{\frac{2}{2(l_1 + 1)}}{\frac{2}{2l_2}} = \frac{l_2}{l_1 + 1}$$

có phân phối Fisher với $2(l_1 + 1)$ và $2l_2$ độ tự do, nên chúng ta có thể dùng trắc nghiệm F để kiểm tra giả thuyết tương đồng $H_0: p = \frac{1}{2}$ hay $H_0: l_1 = l_2$.

Ví dụ 3.5: So sánh kết quả đo vận tốc của 10 ô tô tại cùng một thời điểm của hai loại dụng cụ đo vận tốc được sản xuất ở hai nhà máy khác nhau như sau:

	Ô tô									
	1	2	3	4	5	6	7	8	9	10
$x_{i1}, \text{ km}$	70	85	63	54	65	80	75	95	52	55
$x_{i2}, \text{ km}$	72	86	62	55	63	80	78	90	53	57
$x_{i1} - x_{i2}$	-	-	+	-	+	0	-	+	-	-

$$l_1 = 3, l_2 = 6; F(8, 12) = \frac{6}{4} = 1,5; P > 0,1$$

Vậy giả thuyết $H_0: p = \frac{1}{2}$ được chấp nhận, sự thăng giáng của các nghiệm thức chỉ là sự thăng giáng ngẫu nhiên, nghĩa là kết quả đo vận tốc của hai loại dụng cụ này là như nhau.

3.3.2.2. Các phương pháp so sánh khác

3.3.2.2(a). Dùng trắc nghiệm t

Đây là trắc nghiệm thông dụng nhất, thường được dùng cho trường hợp so sánh hai nghiệm thức với nhau. Để dễ hình dung vấn đề chúng ta vẫn lấy lại số liệu ở ví dụ 3.5 như sau:

	Ô tô										Tổng	Trung bình
	1	2	3	4	5	6	7	8	9	10		
x_{i1}	70	85	63	54	65	80	75	95	52	55	694	69,4
x_{i2}	72	86	62	55	63	80	78	90	53	57		
$x_{i1} - x_{i2}$	-2	-1	1	-1	2	0	-3	5	-1	-2	-2	-0,2

$$n = 10, \sum x_i = (-2) + (-1) + \dots + (-2) = -2, \bar{x} = -0,2$$

$$\sum x_i^2 = (-2)^2 + (-1)^2 + \dots + (-2)^2 = 50$$

$$S_x^2 = \frac{1}{10-1} \left[50 - \frac{(-2)^2}{10} \right] = 5,51 \Rightarrow S_x^2 = \frac{S_x^2}{n} = \frac{5,51}{10} = 0,551$$

$$\text{Vì: } x = x_1 - x_2, \text{ do đó: } S_x^2 = S_{x_1-x_2}^2 = S^2 + S^2 = 2S^2 = 5,51 \Rightarrow S^2 = \frac{5,51}{2} = 2,76$$

là ước lượng tốt nhất của biến lượng quần thể σ^2 .

Giả thuyết $H_0: m (= m_1 - m_2) = 0$

$$t(n-1) = \frac{\bar{x} - m}{S_x} \Rightarrow t(9) = \frac{(-0,2) - 0}{\sqrt{0,551}} = -0,27; P > 0,5$$

Vậy giả thuyết $H_0: m = 0$ được chấp nhận, sự thăng giáng của các nghiệm thức chỉ là sự thăng giáng ngẫu nhiên, nghĩa là kết quả đo vận tốc của hai loại dụng cụ là như nhau:

3.3.2.2.(b). Dùng khoảng tin tưởng

Vẫn lấy lại số liệu ở ví dụ 3.5. Vì rằng:

$$\begin{aligned} t_p(n-1) &< t(n-1) < t_p(n-1) \\ \Rightarrow t_p(n-1) &< \frac{\bar{x} - m}{S_x} < t_p(n-1) \\ \Rightarrow \bar{x} - t_p(n-1) \cdot S_x &< m < \bar{x} + t_p(n-1) \cdot S_x \end{aligned}$$

Với $P = 0,05$ thì $t_p(n-1) = t_{0,05}(10-1) = 2,262$, do đó:

$$-0,2 - 2,262 \cdot \sqrt{0,551} < m < -0,2 + 2,262 \cdot \sqrt{0,551}$$

$\Rightarrow -0,2 - 1,68 < m < -0,2 + 1,68 \Rightarrow -1,88 < m < 1,48$. Vì khoảng tin tưởng của m là $(-1,88; 1,48)$ bao gồm giá trị zero, do đó $m \sim 0 \Rightarrow m_1 - m_2 \sim 0 \Rightarrow m_1 \sim m_2$ (kết luận như trên).

3.3.2.2.(c). Dùng trắc nghiệm F

Vì rằng $t^2(K) = F(1, K)$, do đó, đối với số liệu ở ví dụ 3.5 chúng ta có thể tiến hành phân tích nguồn biến lượng và dùng trắc nghiệm F với giả thuyết $H_0: m_1 = m_2$ như sau:

	Ô tô										Tổng	Trung bình
	1	2	3	4	5	6	7	8	9	10		
X_{i1}	70	85	63	54	65	80	75	95	52	55	694	69,4
X_{i2}	72	86	62	55	63	80	78	90	53	57	696	69,6
Tổng	142	171	125	109	128	160	153	185	105	112	1390	69,5

$$n = 10 + 10 = 20; \sum x_{i1} = 694; \sum x_{i2} = 696; \sum x_i = 70 + 85 + \dots + 53 + 57 = 1390; \sum x_i^2 = 70^2 + 85^2 + \dots + 53^2 + 57^2 = 100074$$

$$Q = 100074 - \frac{1390^2}{20} = 3469$$

$$Q_b = \frac{142^2}{2} + \frac{171^2}{2} + \dots + \frac{112^2}{2} - \frac{1390^2}{20} = 3444$$

$$Q_t = \frac{694^2}{10} + \frac{696^2}{10} - \frac{1390^2}{20} = 0,2$$

$$Q_e = 3469 - (3444 + 0,2) = 24,8 \Rightarrow S^2 = \frac{24,8}{9} = 2,76$$

$$F(1,9) = \frac{0,2}{\frac{1}{24,8}} = 0,073; P > 0,5 \text{ (kết luận như trên).}$$

Để nhận thấy: $t^2(9) = (-0,27)^2 = 0,073 = F(1,9)$ và ước lượng tốt nhất của biến lượng quần thể σ^2 là $S^2 = 2,76$ giống như đã tính ở phần trắc nghiệm t.

3.4. SO SÁNH NHIỀU NGHIỆM THỨC VỚI NHAU

3.4.1. Thuật so sánh đôi chiều

Muốn so sánh nhiều (> 2) nghiệm thức với nhau nhất thiết chúng ta phải tiến hành phân tích nguồn biến động và thực hiện trắc nghiệm F. Nếu như trắc nghiệm F chỉ ra rằng sự thăng giáng của các nghiệm thức chỉ là sự thăng giáng ngẫu nhiên (trường hợp $P > 0,05$), nghĩa là các nghiệm thức là như nhau (giả thuyết H_0 được chấp nhận) thì chúng ta có thể kết luận vấn đề và không cần phải tiến hành các so sánh đôi chiều. Trái lại, nếu như trắc nghiệm F chỉ ra rằng sự thăng giáng của các nghiệm thức là một sự thăng giáng thực sự (trường hợp $P \leq 0,05$) nghĩa là các nghiệm thức là khác nhau (giả thuyết H_0 bị bác bỏ), thì chúng ta chưa thể kết luận vấn đề một cách cụ

thể được, bởi vì ở đây trắc nghiệm chỉ cho ta biết chắc chắn một điều là có ít nhất là 2 nghiệm thức khác nhau trong số các nghiệm thức đem so sánh, chứ không cho chúng ta biết cụ thể là những nghiệm thức nào khác nhau. Trong trường hợp này chúng ta cần tiến hành so sánh đối chiếu giữa hai nghiệm thức, giữa một nghiệm thức và một nhóm nghiệm thức cũng như giữa hai nhóm nghiệm thức với nhau để có được kết quả một cách cụ thể hơn cũng như kết luận rút ra chứa đựng nhiều thông tin hơn. Như đã trình bày ở mục 2.4.2.2, căn bản của phép so sánh đối chiếu là thiết lập khoảng tin tưởng của so sánh C_i và dựa vào khoảng tin tưởng này để kiểm tra giả thuyết tương đồng $H_0^{(i)}$. Nếu khoảng tin tưởng của C_i bao gồm giá trị zero thì giả thuyết tương đồng $H_0^{(i)}$ được chấp nhận (ở mức xác suất sai lầm P), trái lại, nếu như khoảng tin tưởng của C_i không bao gồm giá trị zero thì giả thuyết tương đồng $H_0^{(i)}$ bị bác bỏ (ở mức xác suất sai lầm P). Các mức xác suất sai lầm P thường được xét đến là 0,05; 0,01 và 0,001.

Dạng tổng quát của khoảng tin tưởng của so sánh C_i như sau:

$$C_i = \tilde{C}_i \pm S_{\tilde{C}_i} \cdot T_p^{(a)}(K)$$

Ở đây:

- C_i là so sánh đối chiếu giữa hai nghiệm thức hay nhóm nghiệm thức với giả thuyết tương đồng $H_0^{(i)}$. Chẳng hạn, so sánh đối chiếu thứ nhất được thực hiện là so sánh nghiệm thức I với nghiệm thức II thì:

$$C_1 = m_1 - m_2 \text{ với } H_0^{(1)}: m_1 = m_2$$

C_1 có thể được ký hiệu: $C_{\frac{m_1}{m_2}}$ hay $C_{\frac{\bar{x}_1}{\bar{x}_2}}$ tương tự:

$$C_5 = m_1 - \frac{m_2 + m_3 + m_4}{3} \text{ với } H_0^{(5)}: m_1 = \frac{m_2 + m_3 + m_4}{3}$$

C_5 có thể được ký hiệu: $C_{\frac{m_1}{m_2 m_3 m_4}}$ hay $C_{\frac{\bar{x}_1}{\bar{x}_2 \bar{x}_3 \bar{x}_4}}$

- $\tilde{C}_i$ là ước lượng của C_i chẳng hạn:

$$\text{Nếu } C_1 = m_1 - m_2 \text{ thì } \tilde{C}_1 = \bar{x}_1 - \bar{x}_2$$

$$\text{Nếu } C_1 = m_1 - \frac{m_2 + m_3 + m_4}{3} \text{ thì } \tilde{C}_1 = \bar{x}_1 - \frac{\bar{x}_2 + \bar{x}_3 + \bar{x}_4}{3}$$

- $S_{\tilde{C}_i}$ là độ lệch chuẩn của $\tilde{C}_i$ (biến lượng $\tilde{C}_i$ là $S_{\tilde{C}_i}^2$) chẳng hạn:

$$\text{Nếu } \tilde{C}_i = \bar{x}_1 - \bar{x}_2 \text{ thì } S_{\tilde{C}_i}^2 = \frac{S^2}{n_1} + \frac{S^2}{n_2} = S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \Rightarrow$$

$$\Rightarrow S_{\tilde{C}_i} = \sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

$$\text{Nếu } \tilde{C}_i = \bar{x}_1 - \frac{\bar{x}_2 + \bar{x}_3 + \bar{x}_4}{3} \text{ thì } S_{\tilde{C}_i}^2 = \frac{S^2}{n_1} + \frac{1}{9} \left(\frac{S^2}{n_2} + \frac{S^2}{n_3} + \frac{S^2}{n_4} \right) =$$

$$= S^2 \left[\frac{1}{n_1} + \frac{1}{9} \left(\frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4} \right) \right] \Rightarrow S_{\tilde{C}_i} = \sqrt{S^2 \left[\frac{1}{n_1} + \frac{1}{9} \left(\frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4} \right) \right]}$$

Với $S^2 = \frac{Q_e}{K}$, K là số tự do của Q_e .

- Trong trường hợp của các trắc nghiệm “không xếp hạng các nghiệm thức” thì a là số nghiệm thức, tức $a = l$, còn trong trường hợp của các trắc nghiệm “có xếp hạng các nghiệm thức” thì a là hạng của cặp nghiệm thức đem so sánh, tức $a = 2, 3, \dots, l$. Hạng của cặp nghiệm thức đem so sánh được tính như sau:

Giả sử các nghiệm thức được sắp xếp từ nhỏ đến lớn như sau: $\bar{x}_1 < \bar{x}_2 < \dots < \bar{x}_l$ thì hạng của so sánh $C_{\frac{\bar{x}_i}{\bar{x}_j}}$ là: $a = |i - j| + 1$. Chẳng hạn, với so sánh

$$C_{\frac{\bar{x}_1}{\bar{x}_2}} \text{ thì } a = |1 - 2| + 1 = 2, \text{ với so sánh } C_{\frac{\bar{x}_1}{\bar{x}_l}} \text{ thì } a = |1 - l| + 1 = l.$$

- P là mức xác suất sai lầm được đề cập trong kết luận. P càng nhỏ thì sự khác nhau của cặp nghiệm thức hay nhóm nghiệm thức đem so sánh càng đáng tin tưởng hơn và ngược lại. Ba mức xác suất P thường được xét đến là 0.05, 0.01 và 0.001.

- $T_p^{(a)}(K)$ là đại lượng phụ thuộc vào kiểu trắc nghiệm so sánh đối chiếu được tính từ căn bản của phân phối Student và phân phối Fisher như sau:

(i) Đối với các trắc nghiệm “không xếp hạng” các nghiệm thức:

* Trắc nghiệm LSD:

$$T_p^{(a)}(K) = t_p(K)$$

(Không phụ thuộc vào số lượng nghiệm thức l).

** Trắc nghiệm Scheffe:*

$$T_p^{(a)}(K) = \sqrt{(l-1) \cdot F_p(l-1, K)}$$

(Phụ thuộc vào số lượng nghiệm thức, $a = l$).

** Trắc nghiệm Tukey:*

$$T_p^{(a)}(K) = \frac{q_p^{(a)}(K)}{\sqrt{2}}$$

(Phụ thuộc vào số lượng nghiệm thức, $a = l$).

Ở đây $q_p^{(a)}(K)$ là trị số lấy từ bảng đã được thiết lập sẵn dùng cho trắc nghiệm Tukey (và cả cho trắc nghiệm Newman-Keuls).

** Trắc nghiệm Dunnett:*

Giá trị $T_p^{(a)}(K)$ lấy từ bảng đã được thiết lập sẵn dùng cho trắc nghiệm Dunnett. Đây là trắc nghiệm so sánh đối chứng: với l nghiệm thức trong đó có 1 nghiệm đối chứng thì sẽ có $l - 1$ so sánh đối chứng. ở đây, giá trị $T_p^{(a)}(K)$ phụ thuộc vào số lượng nghiệm thức bao gồm cả nghiệm đối chứng, tức $a = l$.

(ii) Đối với các trắc nghiệm “có xếp hạng” các nghiệm thức:

** Trắc nghiệm Duncan:*

$$T_p^{(a)}(K) = \frac{q_p^{(a)}(K)}{\sqrt{2}}$$

(Phụ thuộc vào hạng của cặp nghiệm thức hay nhóm nghiệm thức đem so sánh đối chiếu, $a = 2, 3, \dots, l$).

Ở đây $q_p^{(a)}(K)$ là trị số lấy từ bảng đã được thiết lập sẵn dùng cho trắc nghiệm Duncan.

** Trắc nghiệm Newman - Keuls:*

$$T_p^{(a)}(K) = \frac{q_p^{(a)}(K)}{\sqrt{2}}$$

(Phụ thuộc vào hạng của cặp nghiệm thức hay nhóm nghiệm thức đem so sánh đối chiếu, $a = 2, 3, \dots, l$).

Ở đây $q_p^{(a)}(K)$ là trị số lấy từ bảng đã được thiết lập sẵn dùng cho trắc nghiệm Newman - Keuls.

Để có một cái nhìn tổng quát hơn về mối quan hệ giữa các trắc nghiệm so sánh đối chiếu, chúng ta thử thiết lập các giá trị $T_p^{(a)}(K)$ với $K = 24$ và $P = 0,05$ ứng với các trắc nghiệm như sau:

Trắc nghiệm	a						
	2	3	4	5	6	7	8
LSD	2,06	2,06	2,06	2,06	2,06	2,06	2,06
Scheffe	2,06	2,61	3,01	3,34	3,62	3,88	4,12
Tukey	2,06	2,50	2,76	2,95	3,09	3,21	3,31
Dunnnett	2,06	2,35	2,53	2,66	2,76	2,84	2,91
Duncan	2,06	2,17	2,23	2,28	2,32	2,34	2,36
Newman - Keuls	2,06	2,50	2,76	2,95	3,09	3,21	3,31

Để nhận thấy là, các trắc nghiệm so sánh đối chiếu khác nhau có giá trị $T_p^{(a)}(K)$ không giống nhau, do đó, trong kết luận ngoài việc đề cập tới mức xác suất sai lầm P , nhất thiết phải nói rõ là thực hiện so sánh với trắc nghiệm nào. Chú ý rằng, trong trường hợp so sánh đối chiếu được thực hiện giữa một nghiệm thức và một nhóm nghiệm thức hoặc giữa hai nhóm nghiệm thức với nhau thì giá trị $T_p^{(a)}(K)$ là giá trị trung bình, chẳng hạn, với so sánh $C_{\frac{m_1 m_2}{m_3}}$

thì $T_p^{(a)}(K) = \frac{T_p^{(a_{13})}(K) + T_p^{(a_{23})}(K)}{2}$, với so sánh $C_{\frac{m_1 m_2}{m_3 m_4}}$ thì:

$$T_p^{(a)}(K) = \frac{T_p^{(a_{13})}(K) + T_p^{(a_{14})}(K) + T_p^{(a_{23})}(K) + T_p^{(a_{24})}(K)}{4},$$

- Đại lượng $S_{C_i} \cdot T_p^{(a)}(K)$ được gọi là “sai biệt nhỏ nhất có ý nghĩa so sánh” (least significant difference, LSD) ở mức xác suất sai lầm P trong trường hợp của các trắc nghiệm “không xếp hạng” các nghiệm thức (trắc nghiệm LSD, Scheffe, Tukey, Dunnnett) và thường được ký hiệu là $D_p^{(a)}(C_i)$, hoặc được gọi là “hạng sai biệt nhỏ nhất có ý nghĩa so sánh”, (least significant range, LSR) ở mức xác suất sai lầm P trong trường hợp của các trắc nghiệm “có xếp hạng” các nghiệm thức (trắc nghiệm Duncan, Newman - Keuls) và thường được ký hiệu là $R_p^{(a)}(C_i)$. Như vậy, khoảng tin tưởng của so sánh C_i có thể được viết dưới dạng sau đây:

$$C_i = \bar{C}_i \pm D_p^{(a)}(C_i) \text{ hoặc } C_i = \bar{C}_i \pm R_p^{(a)}(C_i)$$

Do đó, để nhận thấy là, nếu như $|\bar{C}_i| \geq D_p^{(a)}(C_i)$ (hay $R_p^{(a)}(C_i)$) thì khoảng tin tưởng của C_i không bao gồm giá trị zero nên giả thuyết tương đồng H_0 về sự bằng nhau của cặp nghiệm thức hay nhóm nghiệm thức đem so sánh bị bác bỏ ở mức xác suất sai lầm P , còn nếu như $|\bar{C}_i| < D_p^{(a)}(C_i)$ (hay $R_p^{(a)}(C_i)$) thì khoảng tin tưởng của C_i bao gồm giá trị zero nên giả thuyết tương đồng H_0 về sự bằng nhau của cặp nghiệm thức hay nhóm nghiệm thức đem so sánh được chấp nhận ở mức xác suất sai lầm P . Như thường lệ, mức xác suất $P = 0,05$ là mức xác suất thường được dùng làm cơ sở để bác bỏ giả thuyết H_0 (nếu $P \leq 0,05$) hoặc chấp nhận giả thuyết H_0 (nếu $P > 0,05$). Mức xác suất P càng nhỏ thì sự khác biệt của các nghiệm thức hay nhóm nghiệm thức đem so sánh càng sâu sắc, càng đáng tin tưởng hơn và ngược lại.

3.4.1.1. Trắc nghiệm “không xếp hạng” các nghiệm thức (Multiple test)

3.4.1.1.(a). Trắc nghiệm LSD (LSD test, Multiple test)

Trắc nghiệm LSD là kiểu trắc nghiệm “không xếp hạng” các nghiệm thức hay nhóm nghiệm thức đem so sánh và không phụ thuộc vào số lượng nghiệm thức nhiều hay ít. Giá trị $T_p^{(a)}(K)$ dùng trong trắc nghiệm LSD chính là giá trị $T_p(K)$ của phân phối Student dùng trong trắc nghiệm t . Để minh họa cho kiểu trắc nghiệm này chúng ta lấy lại số liệu của ví dụ 2(27) mà trước đây đã được xử lý bằng trắc nghiệm Scheffe như sau:

$$n_1 = 7; n_2 = 5; n_3 = 3; \bar{x}_1 = 1,43; \bar{x}_2 = 2,4; \bar{x}_3 = 4,0; S^2 = 1,08; K = 12$$

Sau đây là các so sánh đối chiếu được thực hiện:

- So sánh thứ nhất: $C_1 = m_1 - m_2$ với $H_0^{(1)}: m_1 = m_2; P = 0,05$

$$\bar{C}_1 = \bar{x}_1 - \bar{x}_2 = 1,43 - 2,4 = -0,97$$

$$S_{\bar{C}_1}^2 = \frac{S^2}{n_1} + \frac{S^2}{n_2} = S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = 1,08 \left(\frac{1}{7} + \frac{1}{5} \right) = 0,37 \Rightarrow S_{\bar{C}_1} = \sqrt{0,37}$$

$$T_p^{(a)}(K) = T_{0,05}^{(3)}(12) = t_{0,05}(12) = 2,179$$

$$D_p^{(a)}(C_1) = D_{0,05}^{(3)}(C_1) = \sqrt{0,37} \cdot 2,179 = 1,325$$

Vì rằng $|\bar{C}_1| = |-0,97| < D_p^{(a)}(C_1) = 1,325$, do đó giả thuyết $H_0: m_1 = m_2$ được chấp nhận.

Chú ý: có thể dùng trắc nghiệm t như sau:

$$t(12) = \frac{|-0,97|}{\sqrt{0,37}} = 1,59; P > 0,1$$

- So sánh thứ hai: $C_2 = m_1 - m_3$ với $H_0^{(2)}: m_1 - m_3; P = 0,05$

$$\bar{C}_2 = \bar{x}_1 - \bar{x}_3 = 1,43 - 4,0 = -2,57$$

$$S_{\bar{C}_2}^2 = \frac{S^2}{n_1} + \frac{S^2}{n_3} = S^2 \left(\frac{1}{n_1} + \frac{1}{n_3} \right) = 1,08 \left(\frac{1}{7} + \frac{1}{3} \right) = 0,51 \Rightarrow S_{\bar{C}_2} = \sqrt{0,51}$$

$$T_p^{(a)}(K) = 2,179; D_p^{(a)}(C_2) = \sqrt{0,51} \times 2,179 = 1,556; |\bar{C}_2| > D_p^{(a)}(C_2)$$

Giả thuyết $H_0: m_1 - m_3$ bị bác bỏ, nghĩa là $m_1 \neq m_3$. Chú ý có thể dùng trắc nghiệm t như sau:

$$t(12) = \frac{|-2,57|}{\sqrt{0,51}} = 3,6; P < 0,01$$

- So sánh thứ ba: $C_3 = m_2 - m_3$ với $H_0^{(3)}: m_2 = m_3; P = 0,05$

$$\bar{C}_3 = \bar{x}_2 - \bar{x}_3 = 2,4 - 4,0 = -1,6$$

$$S_{\bar{C}_3}^2 = \frac{S^2}{n_2} + \frac{S^2}{n_3} = S^2 \left(\frac{1}{n_2} + \frac{1}{n_3} \right) = 1,08 \left(\frac{1}{5} + \frac{1}{3} \right) = 0,58 \Rightarrow S_{\bar{C}_3} = \sqrt{0,58}$$

$$T_p^{(a)}(K) = 2,179; D_p^{(a)}(C_3) = \sqrt{0,58} \cdot 2,179 = 1,659$$

$|\bar{C}_3| < D_p^{(a)}(C_3)$: Giả thuyết $H_0: m_2 = m_3$ được chấp nhận.

Chú ý: có thể dùng trắc nghiệm t như sau:

$$t(12) = \frac{|-1,6|}{\sqrt{0,58}} = 2,1; P > 0,05$$

Giả sử rằng các nghiệm thức như nhau được biểu thị bởi một gạch liền nhau (-) thì kết quả sẽ được trình bày như sau:

$$\begin{array}{ccc} \bar{x}_1 & \bar{x}_2 & \bar{x}_3 \\ (1,43) & (2,4) & (4,0) \end{array}$$

- So sánh thứ tư:

$$C_4 = \frac{1}{2}(m_1 + m_2) - m_3 \text{ với } H_0^{(4)}: \frac{1}{2}(m_1 + m_2) = m_3; P = 0,05$$

$$\bar{C}_4 = \frac{1}{2}(\bar{x}_1 + \bar{x}_2) - \bar{x}_3 = \frac{1}{2}(1,43 + 2,4) - 4,0 = -2,085$$

$$S_{\bar{C}_4}^2 = \frac{1}{4} \cdot \frac{S^2}{n_1} + \frac{1}{4} \cdot \frac{S^2}{n_2} + \frac{S^2}{n_3} = S^2 \left(\frac{1}{4n_1} + \frac{1}{4n_2} + \frac{1}{n_3} \right) = 1,08 \left(\frac{1}{4,7} + \frac{1}{4,5} + \frac{1}{3} \right) = 0,45$$

$$T_p^{(a)}(K) = 2,179; D_p^{(a)}(C_4) = \sqrt{0,45 \cdot 2,179} = 1,462$$

$$|\bar{C}_4| > D_p^{(a)}(C_4): \text{Giả thuyết } H_0: \frac{1}{2}(m_1 + m_2) = m_3 \text{ bị bác bỏ, nghĩa là}$$

$$\frac{1}{2}(m_1 + m_2) \neq m_3.$$

Chú ý: có thể dùng trắc nghiệm t như sau:

$$t(12) = \frac{-2,085}{\sqrt{0,45}} = 3,1; P < 0,01$$

Chú ý rằng, nếu như số liệu ở các nghiệm thức mà bằng nhau thì công việc tính toán sẽ đơn giản hơn nhiều. Các bạn hãy dùng trắc nghiệm LSD để tiến hành việc so sánh đối chiếu các nghiệm thức và nhóm nghiệm thức ở các ví dụ của mục 3.1.3 cho quen với việc tính toán, đặc biệt là tính toán biến lượng của so sánh ($S_{\bar{C}_i}^2$)

3.4.1.1.(b). Trắc nghiệm Scheffe (Scheffe's test, Scheffe's multiple F test)

Trắc nghiệm Scheffe là kiểu trắc nghiệm “không xếp hạng” các nghiệm thức hay nhóm nghiệm thức đem so sánh nhưng phụ thuộc vào số lượng nghiệm thức l . Giá trị $T_p^{(a)}(K)$ dùng trong trắc nghiệm Scheffe đã được trình bày ở trên. Để minh họa cho kiểu trắc nghiệm này chúng ta lấy lại số liệu của ví dụ ở mục 3.1.3.2 như sau:

$$n_1 = n_2 = n_3 = n_4 = 3 (=r); \bar{x}_1 = 1,07; \bar{x}_2 = 1,22; \bar{x}_3 = 1,25; \bar{x}_4 = 1,29;$$

$$S^2 = 0,00067; K = 6$$

$$\text{- So sánh thứ nhất: } C_1 = m_1 - \frac{m_2 + m_3 + m_4}{3}; H_0^{(1)}: m_1 = \frac{m_2 + m_3 + m_4}{3};$$

$$P = 0,05; \bar{C}_1 = \bar{x}_1 - \frac{\bar{x}_2 + \bar{x}_3 + \bar{x}_4}{3} = 1,07 - \frac{1,22 + 1,25 + 1,29}{3} = -0,183$$

$$S_{\bar{C}_1}^2 = \frac{S^2}{r} + \frac{1}{9} \left(\frac{S^2}{r} + \frac{S^2}{r} + \frac{S^2}{r} \right) = \frac{4S^2}{3r} = \frac{4(0,00067)}{3(3)} = 0,000298$$

$$T_p^{(a)}(K) = T_{0,05}^{(4)}(6) = \sqrt{(4-1) \cdot F_{0,05}(4-1, 6)} = \sqrt{3(4,76)} = 3,779$$

$$D_p^{(a)}(C_1) = \sqrt{0,000298} \cdot 3,779 = 0,065$$

$|\bar{C}_1| > D_p^{(a)}(C_1)$: giả thuyết $H_0: m_1 = \frac{m_2 + m_3 + m_4}{3}$ bị bác bỏ, nghĩa là $m_1 \neq \frac{m_2 + m_3 + m_4}{3}$. Vậy các nghiệm thức nghiên cứu khác với nghiệm chúng (nghiệm thức 1). Dùng trắc nghiệm Dunnett để thực hiện các so sánh đối chứng $\frac{m_1}{m_2}$, $\frac{m_1}{m_3}$ và $\frac{m_1}{m_4}$ (Xem phần dưới).

- So sánh thứ 2: $C_2 = m_2 - m_3$; $H_0^{(2)}: m_2 = m_3$; $P = 0,05$

$$\bar{C}_2 = \bar{x}_2 - \bar{x}_3 = 1,22 - 1,25 = -0,03$$

$$S_{\bar{C}_2}^2 = \frac{S^2}{r} + \frac{S^2}{r} + \frac{2S^2}{r} = \frac{2(0,00067)}{3} = 0,000447; T_p^{(a)}(K) = 3,799$$

$$D_p^{(a)}(C_2) = \sqrt{0,000447} \times 3,779 = 0,0799$$

$|\bar{C}_2| < D_p^{(a)}(C_2)$: giả thuyết $H_0: m_2 = m_3$ được chấp nhận.

- So sánh thứ 3:

$$C_3 = m_2 - m_4; H_0^{(3)}: m_2 = m_3; P = 0,05; \bar{C}_3 = \bar{x}_2 - \bar{x}_4 = 1,22 - 1,29 = -0,07$$

$D_p^{(a)}(C_3) = D_p^{(a)}(C_2) = 0,0799$; $|\bar{C}_3| < D_p^{(a)}(C_3)$: Giả thuyết $H_0: m_2 = m_4$ được chấp nhận. Hơn nữa, giả thuyết $H_0^{(4)}: m_2 = m_3$ cũng được chấp nhận mà không cần phải tính toán để có kết luận. Kết quả so sánh như sau:

$$\begin{array}{ccc} \bar{x}_2 & \bar{x}_3 & \bar{x}_4 \\ (1,22) & (1,25) & (1,29) \end{array}$$

Chú ý rằng, trong trường hợp $a = l = 2$ thì:

$$T_p^{(2)}(K)(LSD) = T_p^{(2)}(K)(Scheffe) = T_p(K)$$

3.4 1.1.(c). Trắc nghiệm Tukey (Tukey's test)

Tương tự như trắc nghiệm Scheffe, trắc nghiệm Tukey là kiểu trắc nghiệm “không xếp hạng” các nghiệm thức hay nhóm nghiệm thức đem so sánh nhưng phụ thuộc vào số lượng nghiệm thức l . Giá trị $T_p^{(a)}(K)$ dùng trong trắc nghiệm Tukey đã được trình bày ở trên. Để minh họa cho kiểu trắc nghiệm này chúng ta lấy lại số liệu của ví dụ 2.(27) như sau:

$$n_1 = 7; n_2 = 5; n_3 = 3; \bar{x}_1 = 1,43; \bar{x}_2 = 2,4; \bar{x}_3 = 4,0; S^2 = 1,08; K = 12$$

- So sánh thứ 1:

$$C_1 = m_1 - m_2; H_0^{(1)}: m_1 = m_2; P = 0,05; \bar{C}_1 = \bar{x}_1 - \bar{x}_2 = 1,43 - 2,4 = 0,97$$

$$S_{\bar{C}_1}^2 = S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = 1,08 \left(\frac{1}{7} + \frac{1}{5} \right) = 0,37$$

$$T_p^{(a)}(K) = T_{0,05}^{(3)}(12) = \frac{q_{0,05}^{(3)}(12)}{\sqrt{2}} = \frac{3,77}{\sqrt{2}} = 2,67$$

$$D_p^{(a)}(C_1) = \sqrt{0,37} \cdot 2,67 = 1,624$$

$|\bar{C}_1| < D_p^{(a)}(C_1)$: giả thuyết $H_0^{(1)}: m_1 = m_2$ được chấp nhận.

- So sánh thứ 2:

$$C_2 = m_1 - m_3; H_0^{(2)}: m_1 = m_3; P = 0,05; \bar{C}_2 = \bar{x}_1 - \bar{x}_3 = 1,43 - 4,0 = -2,57$$

$$S_{\bar{C}_2}^2 = S^2 \left(\frac{1}{n_1} + \frac{1}{n_3} \right) = 1,08 \left(\frac{1}{7} + \frac{1}{3} \right) = 0,51$$

$$T_p^{(a)}(K) = 2,67; D_p^{(a)}(C_2) = \sqrt{0,51} \cdot 2,67 = 1,907$$

$|\bar{C}_2| > D_p^{(a)}(C_2)$: giả thuyết $H_0^{(2)}: m_1 = m_3$ bị bác bỏ, nghĩa là $m_1 \neq m_3$.

- So sánh thứ 3:

$$C_3 = m_2 - m_3; H_0^{(3)}: m_2 = m_3; P = 0,05; \bar{C}_3 = \bar{x}_2 - \bar{x}_3 = 2,4 - 4,0 = -1,6$$

$$S_{\bar{C}_3}^2 = S^2 \left(\frac{1}{n_2} + \frac{1}{n_3} \right) = 1,08 \left(\frac{1}{5} + \frac{1}{3} \right) = 0,58$$

$$T_p^{(a)}(K) = 2,67; D_p^{(a)}(C_3) = \sqrt{0,58} \cdot 2,67 = 2,033$$

$|\bar{C}_3| < D_p^{(a)}(C_3)$: giả thuyết $H_0^{(3)}: m_2 = m_3$ được chấp nhận.

Kết quả: $\frac{m_1 \quad m_2 \quad m_3}{\quad \quad \quad}$

- So sánh thứ 4: $C_4 = \frac{1}{2}(m_1 + m_2) - m_3; H_0^{(4)}: \frac{m_1 + m_2}{2} = m_3; P = 0,05$

$$\bar{C}_4 = \frac{\bar{x}_1 + \bar{x}_2}{2} - \bar{x}_3 = \frac{1,43 + 2,4}{2} - 4,0 = -2,085$$

$$S_{\bar{C}_4}^2 = \frac{1}{4} \left(\frac{S^2}{n_1} + \frac{S^2}{n_2} \right) + \frac{S^2}{n_3} = S^2 \left(\frac{1}{4n_1} + \frac{1}{4n_2} + \frac{1}{n_3} \right) = 1,08 \left(\frac{1}{4 \cdot 7} + \frac{1}{4 \cdot 5} + \frac{1}{3} \right) = 0,45$$

$$T_p^{(a)}(K) = 2,67; D_p^{(a)}(C_4) = \sqrt{0,45} \cdot 2,67 = 1,791$$

$$|\bar{C}_4| > D_p^{(a)}(C_4) : \text{giả thuyết } H_0^{(4)} : \frac{m_1 + m_2}{2} = m_3 \text{ bị bác bỏ, nghĩa là } \frac{m_1 + m_2}{2} \neq m_3.$$

3.4.1.1.(d). Trắc nghiệm Dunnett (Dunnett's test)

Trắc nghiệm Dunnett là kiểu trắc nghiệm “không xếp hạng” các nghiệm thức trong so sánh đối chiếu nhưng phụ thuộc vào số lượng nghiệm thức $a = l$, trong đó có $l - 1$ nghiệm thức nghiên cứu và một nghiệm chứng. Như vậy sẽ có $l - 1$ so sánh đối chứng được thực hiện bằng trắc nghiệm Dunnett đã được trình bày ở trên. Để minh họa cho kiểu trắc nghiệm này chúng ta tiến hành thực hiện so sánh đối chứng cho các nghiệm thức của ví dụ ở mục 3.4.1.1.(b) như sau:

- So sánh đối chứng thứ 1:

$$C_1 = m_1 - m_2; H_0^{(1)} : m_1 = m_2; P = 0,05; \bar{C}_1 = \bar{x}_1 - \bar{x}_2 = 1,07 - 1,22 = -0,15$$

$$S_{\bar{C}_1}^2 = S^2 \left(\frac{1}{r} + \frac{1}{r} \right) = \frac{2S^2}{r} = \frac{2(0,00067)}{3} = 0,000447$$

$$T_p^{(a)}(K) = T_{0,05}^{(4)}(6) = 3,18; D_p^{(a)}(C_1) = \sqrt{0,000447} \cdot 3,18 = 0,067$$

$$|\bar{C}_1| > D_p^{(a)}(C_1) : \text{giả thuyết } H_0^{(1)} : m_1 = m_2 \text{ bị bác bỏ, nghĩa là } m_1 \neq m_2.$$

- Trong trường hợp này các so sánh đối chứng $\frac{m_1}{m_3}, \frac{m_1}{m_4}$ không cần phải thực hiện mà vẫn biết được kết quả là: $m_1 \neq m_3, m_1 \neq m_4$.

Kết quả: $m_1 \quad \underline{m_2} \quad m_3 \quad \underline{m_4}$

3.4.1.2. Trắc nghiệm “có xếp hạng” các nghiệm thức (Multiple range test)

3.4.1.2.(a). Trắc nghiệm Duncan (Duncan's test, Duncan's multiple range test)

Trắc nghiệm Duncan là kiểu trắc nghiệm “có xếp hạng” các nghiệm thức hay nhóm nghiệm thức đem so sánh. Giá trị $T_p^{(a)}(K)$ dùng trong trắc nghiệm Duncan đã được trình bày ở trên. Để minh họa cho kiểu trắc nghiệm này chúng ta lấy lại số liệu của ví dụ 2.(27) như sau:

$$n_1 = 7; n_2 = 5; n_3 = 3; \bar{x}_1 = 1,43; \bar{x}_2 = 2,4; \bar{x}_3 = 4,0; S^2 = 1,08; K = 12$$

- So sánh thứ 1:

$$C_1 = m_1 - m_2; H_0^{(1)}: m_1 = m_2; P = 0,05; \bar{C}_1 = \bar{x}_1 - \bar{x}_2 = 1,43 - 2,4 = -0,97$$

$$S_{\bar{C}_1}^2 = S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = 1,08 \left(\frac{1}{7} + \frac{1}{5} \right) = 0,37$$

$$T_p^{(a)}(K) = T_{0,05}^{(2)}(12) = \frac{-q_{0,05}^{(2)}(12)}{\sqrt{2}} = \frac{3,08}{\sqrt{2}} = 2,178$$

$$D_p^{(a)}(C_1) = \sqrt{0,37} \cdot 2,178 = 1,325$$

$$|\bar{C}_1| < D_p^{(a)}(C_1): \text{giả thuyết } H_0^{(1)}: m_1 = m_2 \text{ được chấp nhận.}$$

- So sánh thứ 2:

$$C_2 = m_1 - m_3; H_0^{(2)}: m_1 = m_3; P = 0,05; \bar{C}_2 = \bar{x}_1 - \bar{x}_3 = 1,43 - 4,0 = -2,57$$

$$S_{\bar{C}_2}^2 = S^2 \left(\frac{1}{n_1} + \frac{1}{n_3} \right) = 1,08 \left(\frac{1}{7} + \frac{1}{3} \right) = 0,51$$

$$T_p^{(a)}(K) = T_{0,05}^{(2)}(12) = \frac{-q_{0,05}^{(2)}(12)}{\sqrt{2}} = \frac{3,23}{\sqrt{2}} = 2,284$$

$$D_p^{(a)}(C_2) = \sqrt{0,51} \cdot 2,284 = 1,631$$

$$|\bar{C}_2| > D_p^{(a)}(C_2): \text{giả thuyết } H_0^{(2)}: m_1 = m_3 \text{ bị bác bỏ, nghĩa là } m_1 \neq m_3.$$

- So sánh thứ 3:

$$C_3 = m_2 - m_3; H_0^{(3)}: m_2 = m_3; P = 0,05;$$

$$\bar{C}_3 = \bar{x}_2 - \bar{x}_3 = 2,4 - 4,0 = -1,6$$

$$S_{\bar{C}_3}^2 = S^2 \left(\frac{1}{n_2} + \frac{1}{n_3} \right) = 1,08 \left(\frac{1}{5} + \frac{1}{3} \right) = 0,58$$

$$T_p^{(a)}(K) = T_{0,05}^{(3)}(12) = \frac{q_{0,05}^{(2)}(12)}{\sqrt{2}} = \frac{3,08}{\sqrt{2}} = 2,178$$

$$D_p^{(a)}(C_3) = \sqrt{0,58} \cdot 2,178 = 1,659$$

$$|\bar{C}_3| < D_p^{(a)}(C_3): \text{giả thuyết } H_0^{(3)}: m_2 = m_3 \text{ được chấp nhận.}$$

Kết quả:

$$\frac{m_1}{m_2} \quad \frac{m_2}{m_3}$$

- So sánh thứ 4:

$$C_4 = \frac{1}{2}(m_1 + m_2) - m_3; H_0^{(4)}: \frac{m_1 + m_2}{2} = m_3; P = 0,05$$

$$\bar{C}_4 = \frac{\bar{x}_1 + \bar{x}_2}{2} - \bar{x}_3 = \frac{1,43 + 2,4}{2} - 4,0 = -2,085$$

$$S_{\bar{C}_4}^2 = \frac{1}{4} \left(\frac{S^2}{n_1} + \frac{S^2}{n_2} \right) + \frac{S^2}{n_3} = \frac{1}{4} \left(\frac{1,08}{7} + \frac{1,08}{5} \right) + \frac{1,08}{3} = 0,45$$

$$T_p^{(a_{13})}(K) = T_{0,05}^{(3)}(12) = 2,284; T_p^{(a_{23})}(K) = T_{0,05}^{(2)}(12) = 2,178$$

$$T_p^{(a)}(K) = \frac{2,284 + 2,178}{2} = 2,231; D_p^{(a)}(C_4) = \sqrt{0,45} \cdot 2,231 = 1,497$$

$$|\bar{C}_4| > D_p^{(a)}(C_4): \text{giả thuyết } H_0^{(4)}: \frac{m_1 + m_2}{2} = m_3 \text{ bị bác bỏ.}$$

$$\text{Nghĩa là: } \frac{m_1 + m_2}{2} \neq m_3.$$

3.4.1.2.(b). Trắc nghiệm Newman - Keuls (Newman - Keuls' test, Newman - Keuls' multiple range test)

Tương tự như trắc nghiệm Duncan, trắc nghiệm Newman - Keuls là kiểu trắc nghiệm có xếp hạng các nghiệm thức hay nhóm nghiệm thức đem so sánh. Giá trị $T_p^{(a)}(K)$ dùng trong trắc nghiệm Newman - Keuls đã được trình bày ở trên. Để tiện so sánh với trắc nghiệm Duncan cũng như các trắc nghiệm so sánh đối chiếu khác chúng ta vẫn lấy lại số liệu ở ví dụ 2.(27) để minh họa cho kiểu trắc nghiệm này.

- So sánh thứ 1:

$$C_1 = m_1 - m_2; H_0^{(1)}: m_1 = m_2; P = 0,05;$$

$$\bar{C}_1 = \bar{x}_1 - \bar{x}_2 = 1,43 - 2,4 = -0,97$$

$$S_{\bar{C}_1}^2 = S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = 1,08 \left(\frac{1}{7} + \frac{1}{5} \right) = 0,37$$

$$T_p^{(a)}(K) = T_{0,05}^{(2)}(12) = \frac{q_{0,05}^{(2)}(12)}{\sqrt{2}} = \frac{3,08}{\sqrt{2}} = 2,178$$

$$D_p^{(a)}(C_1) = \sqrt{0,37} \cdot 2,178 = 1,325$$

$|\bar{C}_1| < D_p^{(a)}(C_1)$: giả thuyết $H_0^{(1)}$: $m_1 = m_2$ được chấp nhận.

- So sánh thứ 2:

$$C_2 = m_1 - m_3; H_0^{(2)}: m_1 = m_3; P = 0,05;$$

$$\bar{C}_2 = \bar{x}_1 - \bar{x}_3 = 1,43 - 4,0 = -2,57$$

$$S_{\bar{C}_2}^2 = S^2 \left(\frac{1}{n_1} + \frac{1}{n_3} \right) = 1,08 \left(\frac{1}{7} + \frac{1}{3} \right) = 0,51$$

$$T_p^{(a)}(K) = T_{0,05}^{(3)}(12) = \frac{q_{0,05}^{(3)}(12)}{\sqrt{2}} = \frac{3,77}{\sqrt{2}} = 2,666$$

$$D_p^{(a)}(C_2) = \sqrt{0,51} \cdot 2,284 = 1,094$$

$|\bar{C}_2| > D_p^{(a)}(C_2)$: giả thuyết $H_0^{(2)}$: $m_1 = m_3$ bị bác bỏ, nghĩa là $m_1 \neq m_3$.

- So sánh thứ 3:

$$C_3 = m_2 - m_3; H_0^{(3)}: m_2 = m_3; P = 0,05;$$

$$\bar{C}_3 = \bar{x}_2 - \bar{x}_3 = 2,4 - 4,0 = -1,6$$

$$S_{\bar{C}_3}^2 = S^2 \left(\frac{1}{n_2} + \frac{1}{n_3} \right) = 1,08 \left(\frac{1}{5} + \frac{1}{3} \right) = 0,58$$

$$T_p^{(a)}(K) = T_{0,05}^{(2)}(12) = \frac{q_{0,05}^{(2)}(12)}{\sqrt{2}} = \frac{3,08}{\sqrt{2}} = 2,178$$

$$D_p^{(a)}(C_3) = \sqrt{0,58} \cdot 2,178 = 1,659$$

$|\bar{C}_3| < D_p^{(a)}(C_3)$: giả thuyết $H_0^{(3)}$: $m_2 = m_3$ được chấp nhận.

Kết quả: $\underline{\quad m_1 \quad m_2 \quad m_3 \quad}$

- So sánh thứ 4:

$$C_4 = \frac{1}{2}(m_1 + m_2) - m_3; H_0^{(4)}: \frac{m_1 + m_2}{2} = m_3; P = 0,05$$

$$\bar{C}_4 = \frac{\bar{x}_1 + \bar{x}_2}{2} - \bar{x}_3 = \frac{1,43 + 2,4}{2} - 4,0 = -2,085$$

$$S_{\bar{C}_4}^2 = \frac{1}{4} \left(\frac{S^2}{n_1} + \frac{S^2}{n_2} \right) + \frac{S^2}{n_3} = \frac{1}{4} \left(\frac{1,08}{7} + \frac{1,08}{5} \right) + \frac{1,08}{3} = 0,45$$

$$T_p^{(a_{13})}(K) = T_{0,05}^{(3)}(12) = 2,666; \quad T_p^{(a_{23})}(K) = T_{0,05}^{(2)}(12) = 2,178$$

$$T_p^{(a)}(K) = \frac{2,666 + 2,178}{2} = 2,422; \quad D_p^{(a)}(C_4) = \sqrt{0,45 \cdot 2,422} = 1,025$$

$$|\bar{C}_4| > D_p^{(a)}(C_4) : \text{giả thuyết } H_0^{(4)} : \frac{m_1 + m_2}{2} = m_3 \text{ bị bác bỏ.}$$

$$\text{Nghĩa là } \frac{m_1 + m_2}{2} \neq m_3.$$

3.4.2. Thuật so sánh “phân định”

So sánh “phân định” là một kiểu so sánh đối chiếu trong đó các nghiệm thức được phân thành từng “cặp” hay từng “nhóm cặp” để so sánh với nhau và mỗi “cặp” hay mỗi “nhóm cặp” như vậy sẽ ứng với 1 độ tự do.

Có hai kiểu so sánh phân định là *so sánh “phân định độc lập”* (independent comparison, orthogonal comparison) và *so sánh “phân định phi độc lập”* (nonindependent comparison, nonorthogonal comparison). Để minh định trước hết cần phải biết thế nào là một so sánh.

3.4.2.1. Thế nào là một so sánh

Giả sử rằng nghiệm thức i có n_i số liệu, với tổng giá trị của các số liệu ở nghiệm thức đó là T_i và trung bình nghiệm thức là $\bar{x}_i$.

- Đại lượng $C = \sum \lambda_i T_i$ được coi là một so sánh nếu như $\sum n_i \lambda_i = 0$. Trong trường hợp số số liệu ở mỗi nghiệm thức đều bằng nhau và bằng r thì $\sum n_i \lambda_i = r \sum \lambda_i = 0 \Rightarrow \sum \lambda_i = 0$. Để minh họa chúng ta lấy lại số liệu của ví dụ đã cho ở mục 3.1.3.1 như sau:

Nghiệm thức i	Số liệu	Tổng T_i	Số số liệu n_i	Trung bình $\bar{x}_i$
1	57,8 ; 56,2; 61,9; 54,4; 53,6; 56,4; 53,2	393,5	7	56,21
2	64,2; 58,7; 63,1; 62,5; 59,8; 59,2	367,5	6	61,25
3	62,5; 65,2; 67,6; 69,9; 69,4; 70,1; 67,8; 67,0; 68,5; 62,4	670,4	10	67,04
Tổng		1431,4	23	$\bar{x} = 62,23$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Nghiệm thức	2	490,4	
Sai lệch ngẫu nhiên	20	152,3	
Tổng	22	642,7	

+ So sánh nghiệm thức thứ 1 với nghiệm thức thứ 2 là:

$$C_{\frac{T_1}{T_2}} = 6T_1 - 7T_2; \lambda_1 = 6; \lambda_2 = -7 \text{ (vì } n_1\lambda_1 + n_2\lambda_2 = 7(6) + 6(-7) = 0)$$

+ So sánh nghiệm thức thứ 1 với nghiệm thức thứ 3 là:

$$C_{\frac{T_1}{T_3}} = 10T_1 - 7T_3; \lambda_1 = 10; \lambda_3 = -7 \text{ (vì } n_1\lambda_1 + n_3\lambda_3 = 7(10) + 10(-7) = 0)$$

+ So sánh nghiệm thức thứ 2 với nghiệm thức thứ 3 là:

$$C_{\frac{T_2}{T_3}} = 10T_2 - 6T_3; \lambda_2 = 10; \lambda_3 = -6 \text{ (vì } n_2\lambda_2 + n_3\lambda_3 = 6(10) + 10(-6) = 0)$$

So sánh nghiệm thức thứ 1 với nhóm gồm nghiệm thức 2 và 3 là:

$$C_{\frac{T_1}{T_2 T_3}} = 16T_1 - 7(T_2 + T_3) = 16T_1 - 7T_2 - 7T_3; \lambda_1 = 16; \lambda_2 = -7; \lambda_3 = -7$$

$$\text{(vì } n_1\lambda_1 + n_2\lambda_2 + n_3\lambda_3 = 7(16) + 6(-7) + 10(-7) = 0)$$

- Nếu $C = \sum \lambda_i T_i$ là một so sánh thì đại lượng $Q_i(C) = \frac{C^2}{\sum n_i \lambda_i^2}$ là tổng các

bình phương của so sánh ứng với 1 độ tự do, tức biến lượng của so sánh ($S_C^2 = \frac{Q_i(C)}{1} = Q_i(C)$). Trong trường hợp số số liệu ở các nghiệm thức đều bằng nhau và bằng r thì:

$$S_C^2 = \frac{Q_i(C)}{1} = Q_i(C) = \frac{C^2}{\sum n_i \lambda_i^2} = \frac{C^2}{r \sum \lambda_i^2} = \frac{C^2}{r \sum \lambda_i^2}$$

+ Tổng các bình phương của so sánh (tức biến lượng của so sánh) thứ 1 giữa nghiệm thức thứ 1 và nghiệm thức 2 như sau:

$$C_1 = 6T_1 - 7T_2 = 6(393,5) - 7(367,5) = -211,5$$

$$Q_i(C_1) = S_{C_1}^2 = \frac{C_1^2}{n_1 \lambda_1^2 + n_2 \lambda_2^2} = \frac{(-211,5)^2}{7(6)^2 + 6(-7)^2} = 81,9$$

Thực ra, $Q_i(C_1)$ có thể được tính theo cách đã trình bày ở mục phân tích nguồn biến lượng như sau:

$$Q_i(C_1) = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} - \frac{(T_1 + T_2)^2}{n_1 + n_2} = \frac{393,5^2}{7} + \frac{367,5^2}{6} - \frac{(393,5 + 367,5)^2}{7+6} = 81,9$$

+ Tương tự đối với so sánh thứ 2 như sau:

$$C_2 = 10T_1 - 7T_3 = 10(393,5) - 7(670,4) = -757,8$$

$$Q_i(C_2)(=S_{C_2}^2) = \frac{C_2^2}{n_1\lambda_1^2 + n_3\lambda_3^2} = \frac{(-757,8)^2}{7(10)^2 + 10(-7)^2} = 482,6$$

Hoặc có thể tính theo cách khác như sau:

$$Q_i(C_2) = \frac{T_1^2}{n_1} + \frac{T_3^2}{n_3} - \frac{(T_1 + T_3)^2}{n_1 + n_3} = \frac{393,5^2}{7} + \frac{670,4^2}{10} - \frac{(393,5 + 670,4)^2}{7+10} = 482,6$$

+ Đối với so sánh thứ 3 như sau:

$$C_3 = 10T_2 - 6T_3 = 10(367,5) - 6(670,4) = -347,4$$

$$Q_i(C_3)(=S_{C_3}^2) = \frac{C_3^2}{n_2\lambda_2^2 + n_3\lambda_3^2} = \frac{(-347,4)^2}{6(10)^2 + 10(-6)^2} = 125,7$$

Hoặc có thể tính theo cách khác như sau:

$$Q_i(C_3) = \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} - \frac{(T_2 + T_3)^2}{n_2 + n_3} = \frac{367,5^2}{6} + \frac{670,4^2}{10} - \frac{(367,5 + 670,4)^2}{6+10} = 125,7$$

+ Đối với so sánh thứ 4 như sau:

$$C_4 = 16T_1 - 7T_2 - 7T_3 = 16(393,5) - (367,5) - 7(670,4) = -969,3$$

$$Q_i(C_4)(=S_{C_4}^2) = \frac{C_4^2}{n_1\lambda_1^2 + n_2\lambda_2^2 + n_3\lambda_3^2} = \frac{(-969,3)^2}{7(16)^2 + 6(-7)^2 + 10(-7)^2} = 364,7$$

Hoặc có thể tính theo cách khác như sau:

$$Q_i(C_4) = \frac{T_1^2}{n_1} + \frac{(T_2 + T_3)^2}{n_2 + n_3} - \frac{(T_1 + T_2 + T_3)^2}{n_1 + n_2 + n_3} = \frac{393,5^2}{7} + \frac{(367,5 + 670,4)^2}{6+10} - \frac{1431,4^2}{23} = 364,7$$

- Vì rằng $t(K) = \sqrt{F(1, K)}$, do đó, đối với các so sánh đối chiếu bằng cách “phân định” tổng các bình phương so sánh thành một độ tự do chúng ta có thể dùng trực nghiệm F hay trực nghiệm t đều được cả.

- Đối với so sánh thứ 1 $\left(\frac{T_1}{T_2} \right)$: $F(1, 20) = \frac{S_{C_1}^2}{S^2} = \frac{81,9}{7,6} = 10,78$; $P < 0,005$

Hoặc: $t(20) = \frac{S_{C_1}}{S} = \frac{\sqrt{81,9}}{\sqrt{7,6}} = 3,28$; $P < 0,005$; (Hoặc $t(20) = \sqrt{F(1, 20)} = \sqrt{10,78} = 3,28$).

Vậy giả thuyết tương đồng $H_0^{(1)}$: $m_1 = m_2$ bị bác bỏ, nghĩa là $m_1 \neq m_2$.

+ Vì rằng $S_{C_1}^2 < S_{C_2}^2, S_{C_3}^2, S_{C_4}^2$, do đó, không cần phải thực hiện các so sánh còn lại chúng ta vẫn có được kết quả so sánh là: $m_1 \neq m_3, m_2 \neq m_3, m_1 \neq \frac{m_2 + m_3}{2}$.

- Bản chất của so sánh “phân định” là trực nghiệm LSD (LSD test), tức trực nghiệm t trong so sánh đối chiếu (multiple t test). Thực vậy, chúng ta có thể so sánh đối chiếu nghiệm thức thứ 1 và 2 bằng trực nghiệm t, với giả thuyết H_0 : $m_1 = m_2$, như sau:

$$\begin{aligned} \bar{C} &= \bar{x}_1 - \bar{x}_2 = \frac{T_1}{n_1} - \frac{T_2}{n_2}; S_C^2 = S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = S^2 \cdot \frac{n_1 + n_2}{n_1 n_2} \\ t(K) &= \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{S^2 \cdot \frac{n_1 + n_2}{n_1 n_2}}} = \frac{\left(\frac{T_1}{n_1} - \frac{T_2}{n_2} \right) - 0}{\sqrt{S^2 \cdot \frac{n_1 + n_2}{n_1 n_2}}} \Rightarrow t^2(K) = \frac{\left(\frac{T_1}{n_1} - \frac{T_2}{n_2} \right)^2}{S^2 \cdot \frac{n_1 + n_2}{n_1 n_2}} = \\ &= \frac{(n_2 T_1 - n_1 T_2)^2}{S^2} \cdot \frac{1}{(n_1 n_2)^2} \cdot \frac{n_1 n_2}{n_1 + n_2} = \frac{(n_2 T_1 - n_1 T_2)^2}{S^2} \cdot \frac{1}{n_1 n_2 (n_1 + n_2)} = \\ &= \frac{(n_2 T_1 - n_1 T_2)^2}{S^2} \cdot \frac{1}{n_1 (n_2)^2 + n_2 (-n_1)^2} = \frac{S_{C_1}^2}{S^2} = F(1, K) \end{aligned}$$

$$\begin{aligned}
 (\forall C_1 = n_2 T_1 - n_1 T_1 \Rightarrow Q_1(C_1) (= S_{C_1}^2) &= \frac{C_1^2}{n_1(n_2)^2 + n_2(-n_1)^2} = \\
 &= \frac{(n_2 T_1 - n_1 T_2)^2}{n_1(n_2)^2 + n_2(-n_1)^2}.
 \end{aligned}$$

Cụ thể từ số liệu của ví dụ trên như sau:

$$\bar{C} = \bar{x}_1 - \bar{x}_2 = 56,21 - 61,25 = -5,04$$

$$S_{\bar{C}}^2 = S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = 7,6 \left(\frac{1}{7} + \frac{1}{6} \right) = 2,35$$

$$t(20) = \frac{|-5,04|}{\sqrt{2,35}} = 3,28; P < 0,005 \text{ (như trên)}$$

3.4.2.2. So sánh “phân định độc lập” và so sánh “phân định phi độc lập”

- Hai so sánh “phân định” $C_1 = \sum \lambda_i^{(1)} T_i$ và $C_2 = \sum \lambda_i^{(2)} T_i$ là “độc lập” nếu $\sum n_i \lambda_i^{(1)} \lambda_i^{(2)} = 0$, và là “phi độc lập” nếu: $\sum n_i \lambda_i^{(1)} \lambda_i^{(2)} \neq 0$. Trong trường hợp số số liệu ở các nghiệm thức đều bằng nhau và bằng r thì:

$\sum n_i \lambda_i^{(1)} \lambda_i^{(2)} = \sum r \lambda_i^{(1)} \lambda_i^{(2)} = r \sum \lambda_i^{(1)} \lambda_i^{(2)} = 0 \Rightarrow \sum \lambda_i^{(1)} \lambda_i^{(2)} = 0$. Để minh họa chúng ta vẫn lấy số liệu của ví dụ đã cho ở trên như sau:

Xét hai so sánh: $C_{\frac{T_1}{T_2 T_3}}$ và $C_{\frac{T_2}{T_3}}$:

Tổng nghiệm thức, T_i	T_1	T_2	T_3
Số số liệu, n_i	7	6	10
$C_{\frac{T_1}{T_2 T_3}}$	16 T_1	- 7 T_2	- 7 T_3
$C_{\frac{T_2}{T_3}}$	0. T_1 +	10 T_2	- 6 T_3

Vì rằng $\sum n_i \lambda_i^{(1)} \lambda_i^{(2)} = 7(16)(0) + 6(-7)(10) + 10(-7)(-6) = 0$, do đó, các so sánh $C_{\frac{T_1}{T_2 T_3}}$ và $C_{\frac{T_2}{T_3}}$ là các so sánh “phân định độc lập”.

- Xét 2 so sánh: $\frac{C_{T_1}}{T_2}$ và $\frac{C_{T_1}}{T_3}$

Tổng nghiệm thức, T_i	T_1	T_2	T_3
Số số liệu, n_i	7	6	10
$\frac{C_{T_1}}{T_2}$	$6T_1$	$-7T_2$	$-0.T_3$
$\frac{C_{T_1}}{T_3}$	$10T_1$	$+0.T_2$	$-7T_3$

Vì rằng $\sum n_i \lambda_i^{(1)} \lambda_i^{(2)} = 7(6)(10) + 6(-7)(0) + 10(0)(-7) = 420$, do đó, các so sánh $\frac{C_{T_1}}{T_2}$ và $\frac{C_{T_1}}{T_3}$ là các so sánh “phân định độc lập”.

- Đối với kiểu so sánh “phân định độc lập”, từ l nghiệm thức chúng ta có thể phân lập thành một bộ gồm $l - 1$ so sánh “độc lập” nhau và khi đó thì:

$$\sum_{j=1}^{l-1} \frac{C_j^2}{\sum_{i=1}^l n_i \lambda_i^{(j)2}} = \sum_{j=1}^{l-1} Q_i(C_j) = Q_l \text{ nghĩa là, từ tổng các bình phương của}$$

nghiệm thức Q_l với $l - 1$ độ tự do chúng ta đã “phân định” ra thành $l - 1$ phần, mỗi phần ứng với một độ tự do, một cách “độc lập” với nhau. Chẳng hạn, ở ví dụ đã cho ở trên, từ 3 nghiệm thức chúng ta có thể phân lập thành một bộ gồm 2 so sánh “độc lập” với nhau $\frac{C_{T_1}}{T_2 T_3}$ và $\frac{C_{T_2}}{T_3}$ như sau:

$$\begin{array}{ccc} Q_{\frac{T_1}{T_2 T_3}} & + & Q_{\frac{T_2}{T_3}} = Q_l \\ 364,7 & & 125,7 & & 490,4 \\ (1 \text{ độ tự do}) & & (1 \text{ độ tự do}) & & (2 \text{ độ tự do}) \end{array}$$

Thuật so sánh “phân định” giúp chúng ta đánh giá kết quả nghiên cứu của mình “tinh tế” hơn, nhận được nhiều thông tin hơn, đặc biệt là đối với các nghiên cứu về sự tác động phối hợp đa yếu tố.

Ví dụ 3.6: Đánh giá tác dụng tăng lượng (g) của khẩu phần thức ăn chính (protein cao/thấp) và bổ sung (thịt bò, ngũ cốc, thịt heo) đối với các chuột đực cùng lứa.

Protein mức cao			Protein mức thấp		
Thịt bò	Ngũ cốc	Thịt heo	Thịt bò	Ngũ cốc	Thịt heo
73	98	94	90	107	49
102	74	79	76	95	82
118	56	96	90	97	73
104	111	98	64	80	86
81	95	102	86	98	81
107	88	102	51	74	97
100	82	108	72	74	106
87	77	91	90	67	70
117	86	120	95	89	61
111	92	105	78	58	82
1000	859	995	792	839	787

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Nghiệm thức	5	4612,9	$S^2 = 214,6$
Sai lệch ngẫu nhiên	54	11586,0	
Tổng	59	16198,9	
$F(5, 54) = \frac{\frac{4612,9}{5}}{\frac{11586,0}{54}} = 4,3; P < 0,005$			

Bây giờ, nếu như chúng ta tiến hành “phân định” cho 6 nghiệm thức trên thành một bộ gồm 5 so sánh “độc lập” với nhau như sau đây:

Ví dụ 3.7: Đánh giá tác dụng tăng lượng (kg) của khẩu phần thức ăn chính (chứa 12% và 14% protein) và bổ sung (chứa 0% và 0,6% Lysine) đối với các heo đực và cái khác lứa.

Lysine	Protein	Phái tính	Lứa								Tổng
			1	2	3	4	5	6	7	8	
0	12	đực	1,11	0,97	1,09	0,99	0,85	1,21	1,29	0,96	8,47
		cái	1,03	0,97	0,99	0,99	0,99	1,21	1,19	1,24	-
0,6	14	đực	1,52	1,45	1,27	1,22	1,67	1,24	1,34	1,32	8,61
		cái	1,48	1,22	1,53	1,19	1,16	1,57	1,13	1,43	11,03
	12	đực	1,22	1,13	1,34	1,41	1,34	1,19	1,25	1,32	10,71
		cái	0,87	1,00	1,16	1,29	1,00	1,14	1,36	1,32	10,20
	14	đực	1,38	1,08	1,40	1,21	1,46	1,39	1,17	1,21	9,14
		cái	1,09	1,09	1,47	1,43	1,24	1,17	1,01	1,13	10,30
											9,63
	Tổng			9,70	8,91	10,25	9,73	9,71	10,12	9,74	9,93

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Lứa (khối)	7	0,1411	S ² = 0,0224
Nghiem thức	7	0,7986	
Sai lệch ngẫu nhiên	49	1,0994	
Tổng	63	2,0391	

$$F(7, 49) = \frac{\frac{0,7986}{7}}{\frac{1,0994}{49}} = 5,08; P < 0,005$$

Bây giờ, chúng ta tiến hành “phân định” cho 8 nghiệm thức trên thành một bộ gồm 7 so sánh “độc lập” với nhau như sau:

So sánh	Lysine = 0				Lysine = 0,6%				C _j	$\Sigma n_i \lambda_i^{(j)^2}$	$\frac{C_j^2}{\Sigma n_i \lambda_i^{(j)^2}}$	F _j	P _j
	Prot = 12%		Prot = 14%		Prot = 12%		Prot = 14%						
	Đực 8,47	Cái 8,61	Đực 11,03	Cái 10,71	Đực 10,20	Cái 9,14	Đực 10,3 0	Cái 9,63					
1.Đực/Cái	-1	+1	-1	+1	-1	+1	-1	+1	-1,91	8(8)	0,0570	2,54	> 0,1
2. Protein 12/14%	-1	-1	+1	+1	-1	-1	+1	+1	5,25	8(8)	0,4307	19,23	< 0,001
3. Phối tính. Protein	+1	-1	-1	+1	+1	-1	-1	+1	-0,07	8(8)	0,0001	0,005	> 0,5
4. Lysine 0/0.6%	-1	-1	-1	-1	+1	+1	+1	+1	0,45	8(8)	0,0032	0,143	> 0,5
5. Phối tính. Lysine	+1	-1	+1	-1	-1	+1	-1	+1	-1,55	8(8)	0,0375	1,67	> 0,2
6. Protein. Lysine	+1	+1	-1	-1	-1	-1	+1	+1	-4,07	8(8)	0,2588	11,55	< 0,005
7. Phối tính. Prot Lys	-1	+1	+1	-1	+1	-1	-1	+1	0,85	8(8)	0,0113	0,505	> 0,4
Tổng											0,7986		

Ví dụ 3.8: Đánh giá tác dụng tăng lượng (kg) của Lysine (L), Methionine (M) và Protein (P) đối với các heo cái khác lứa.

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Lứa (Khối)	1	0,0951	$S^2 = 0,0214$
Nghiệm thức	17	1,0269	
Sai lệch ngẫu nhiên	17	0,3633	
Tổng	35	1,4853	
$F(17,17) = \frac{\frac{1,0269}{17}}{\frac{0,3633}{17}} = 2,83; P < 0,025$			

Chú ý rằng, ngoài việc “phân định” cho tổng các bình phương của nghiệm thức Q_t ứng với mỗi so sánh C_j là $Q_t(C_j)$, chúng ta còn có thể “phân định” cho tổng các bình phương của sai lệch ngẫu nhiên Q_e ứng với mỗi so sánh C_j là $Q_e(C_j)$. Như vậy, từ l nghiệm thức chúng ta có thể thiết lập một bộ gồm $l - 1$ so sánh “độc lập” với nhau:

$$Q_t = \sum_{j=1}^{l-1} Q_t(C_j) \quad \text{và} \quad Q_e = \sum_{j=1}^{l-1} Q_e(C_j)$$

Ví dụ 3.9: Đánh giá tác dụng bảo quản hạt giống của các loại thuốc diệt nấm : A (nghiệm chứng, không xử lý thuốc) B và C (dùng các loại thuốc diệt nấm thuộc nhóm thủy ngân), D và H (dùng các loại thuốc diệt nấm thuộc nhóm phi thủy ngân của công ty I), E, F và G (dùng các loại thuốc diệt nấm thuộc nhóm phi thủy ngân của công ty II, trong đó F và G là thuốc mới).

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Khối	5	102,50	$S^2 = 5,78$
Nghiệm thức	7	1210,57	
Sai lệch ngẫu nhiên	35	202,18	
Tổng	47	1515,25	
$F(7,38) = \frac{\frac{1,210,57}{7}}{\frac{202,18}{35}} = 29,94; P < 0,005$			

Lựa	L = 0				L ₁ = 0,05%			
	M ₀ = 0		M ₁ = 0,025%		M ₂ = 0,050%		M ₃ = 0,050%	
	P ₁ = 12%	P ₂ = 14%	P ₁	P ₂	P ₁	P ₂	P ₁	P ₂
1	1,11	1,52	1,09	1,27	0,85	1,67	1,03	1,24
2	0,97	1,45	0,99	1,22	1,21	1,24	1,21	1,34
Tổng	2,08	2,97	2,08	2,49	2,06	2,91	2,24	2,58
So sánh								
1. P ₁ /P ₂	-1	+1	-1	+1	-1	+1	-1	+1
2. M ₀ /M ₁ M ₂	-2	-2	+1	+1	+1	+1	+1	+1
3. M ₁ /M ₂	0	0	-1	-1	+1	+1	-1	-1
4. P.M ₀ /M ₁ M ₂	+2	-2	-1	+1	-1	+1	-1	+1
5. P.M ₁ /M ₂	0	0	+1	-1	-1	+1	+1	-1
P.M								
6. L ₁ /L ₁ L ₂	-2	-2	-2	-2	-2	-2	+1	+1
7. L ₁ /L ₂	0	0	0	0	0	0	-1	-1
8. P.L ₁ /L ₁ L ₂	+2	-2	+2	-2	+2	-2	-1	+1
9. P.L ₁ /L ₂	0	0	0	0	0	0	+1	-1
P.L								
10. (M ₀ /M ₁ M ₂)(L ₁ /L ₂)	+4	+4	-2	-2	-2	-2	+1	+1
11. (M ₀ /M ₁ M ₂)(L ₁ /L ₂)P	0	0	0	0	0	0	-1	-1
12. (M ₁ /M ₂)(L ₁ /L ₂)	0	0	+2	+2	-2	-2	-1	+1
13. (M ₁ /M ₂)(L ₁ /L ₂)P	0	0	0	0	0	0	+1	-1
M.L								
14. (M ₀ /M ₁ M ₂)(L ₁ /L ₂)P	+4	+4	+2	-2	+2	-2	-1	+1
15. (M ₀ /M ₁ M ₂)(L ₁ /L ₂)P	0	0	0	0	0	0	+1	-1
16. (M ₁ /M ₂)(L ₁ /L ₂)P	0	0	-2	+2	+2	-2	+1	-1
17. (M ₁ /M ₂)(L ₁ /L ₂)P	0	0	0	0	0	0	-1	+1
M.L.P								
Tổng								

L ₂ = 0,10%													
		M ₁ = 0		M ₁ = 0,025%		M ₂ = 0,050%							
Lúa	Số sánh	P ₁	P ₂	P ₁	P ₂	P ₁	P ₂	Tống	C ₁	Σ n ₁ λ ₁ (j) ²	C ₁ ² / Σ n ₁ λ ₁ (j) ²	F ₁	P ₁
		1,22	1,38	1,34	1,40	1,34	1,46						
1		1,13	1,08	1,41	1,21	1,19	1,39	21,80					
2													
	Tống	2,35	2,46	2,75	2,62	2,53	2,85	45,45					
1. P ₁ /P ₂		-1	+1	-1	+1	-1	+1	4,51		2(18)	0,5650	26,40	<0,001
2. M ₂ /M ₁ M ₂		-2	-2	+1	+1	+1	+1	-0,27		2(38)	0,0010	0,05	>0,5
3. M ₁ /M ₂		0	0	-1	-1	+1	+1	0,71		2(12)	0,0210	0,98	>0,3
4. P ₁ M ₁ /M ₁ M ₂		+2	-2	-1	+1	-1	+1	-0,83		2(36)	0,0096	0,45	>0,5
5. P ₁ M ₁ /M ₂		0	0	+1	-1	-1	-1	1,51		2(12)	0,0950	4,44	>0,05
P.M													
6. L ₂ /L ₁ L ₂		+1	+1	+1	+1	+1	+1	1,68		2(36)	0,0392	1,83	>0,1
7. L ₁ /L ₂		+1	+1	+1	+1	+1	+1	0,24		2(12)	0,0024	0,11	>0,5
8. P ₁ L ₁ /L ₁ L ₂		-1	+1	-1	+1	-1	+1	-1,94		2(36)	0,0523	2,44	>0,1
9. P ₁ L ₁ /L ₂		-1	+1	-1	+1	-1	+1	-1,78		2(12)	0,1320	6,17	<0,025
P.L													
10. (M ₂ /M ₁ M ₂)(L ₂ /L ₁ L ₂)		-2	-2	+1	+1	+1	+1	1,41		2(72)	0,0138	0,65	>0,4
11. (M ₂ /M ₁ M ₂)(L ₂ /L ₂)		-2	-2	+1	+1	+1	+1	1,95		2(24)	0,0792	3,70	>0,05
12. (M ₁ /M ₂)(L ₂ /L ₁ L ₂)		0	0	-1	-1	+1	+1	-0,49		2(24)	0,0050	0,23	>0,5
13. (M ₁ /M ₂)(L ₂ /L ₂)		0	0	-1	-1	+1	+1	-0,27		2(8)	0,0046	0,22	>0,5
M.L											0,1026	1,20	>0,1
14. (M ₂ /M ₁ M ₂)(L ₂ /L ₁ L ₂)P		+2	-2	-1	+1	-1	+1	0,73		2(72)	0,0037	0,17	>0,5
15. (M ₂ /M ₁ M ₂)(L ₂ /L ₂)P		+2	-2	-1	+1	-1	+1	0,23		2(24)	0,0010	0,05	>0,5
16. (M ₁ /M ₂)(L ₂ /L ₁ L ₂)P		0	0	+1	-1	-1	+1	0,19		2(24)	0,0007	0,03	>0,5
17. (M ₁ /M ₂)(L ₂ /L ₂)P		0	0	+1	-1	-1	+1	-0,15		2(8)	0,0014	0,07	>0,5
M.L.P											0,0068	0,08	>0,1
Tống											1,0261		

Khối	Thử nghiệm								Tổng					
	A	B	c	D	E	F	G	H						
1	8	16	14	10	8	8	7	12	83					
2	8	19	16	11	7	8	6	19	94					
3	9	24	14	12	1	3	6	9	78					
4	7	22	13	8	1	3	6	11	71					
5	7	19	14	7	3	3	4	9	66					
6	5	19	13	3	2	7	4	5	58					
Tổng So sánh	44	119	84	51	22	32	33	65	450 C_j	$\sum n_i \lambda_i^{(j)^2}$	$\frac{C_i^2}{\sum n_i \lambda_i^{(j)^2}}$	F _j	P _j	
1. A/BCDEFGH	-7	+1	+1	+1	+1	+1	+1	+1	98	6(56)	28,58	4,94	< 0,05	
2. BC/DEFGH	0	+5	+5	-2	-2	-2	-2	-2	609	6(70)	883,05	152,78	< 0,001	
3. B/C	0	+1	-1	0	0	0	0	0	35	6(2)	102,08	17,66	< 0,001	
4. DH/EFG	0	0	0	+3	-2	-2	-2	+3	174	6(30)	168,20	29,10	< 0,001	
5. D/H	0	0	0	+1	0	0	0	-1	-14	6(2)	16,33	2,82	> 0,1	
6. E/FG	0	0	0	0	+2	-1	-1	0	-21	6(6)	12,25	2,12	> 0,1	
7. F/G	0	0	0	0	0	+1	-1	0	-1	6(2)	0,08	0,014	> 0,5	
Tổng											1210,57			

Bây giờ, chúng ta tiến hành “phân định” cho tổng các bình phương của sai lệch ngẫu nhiên $Q_e = 202,18$ với 35 độ tự do ra thành 7 “phần” ứng với 7 so sánh, mỗi “phần” là 5 độ tự do như sau:

Trước hết, cần tính “lượng so sánh” ở từng “khối”, chẳng hạn như:

- Lượng so sánh 1 ở khối 1 là:

$$C_{1[1]} = -7(8) + 1(16 + 14 + 10 + 8 + 8 + 7 + 12) = 19$$

Với $\Sigma \lambda_i^{(1)2} = (-7)^2 + 7(+1)^2 = 56$

.....

- Lượng so sánh 7 ở khối 6 là:

$$C_{7[6]} = 1(7) + (-1)(4) = 3$$

Với $\Sigma \lambda_i^{(7)2} = (+1)^2 + 7(-1)^2 = 2$

Kết quả của tính toán “lượng so sánh” ở từng “khối” được trình bày ở bảng sau đây:

Khối	So sánh						
	1	2	3	4	5	6	7
1	19	60	2	20	-2	1	1
2	30	73	3	48	-8	0	2
3	6	128	10	43	3	-7	-3
4	15	117	9	37	-3	-7	-3
5	10	113	5	28	-2	-1	-1
6	18	118	6	-2	-2	-7	3
Tổng, C_i	98	609	35	174	-14	-21	-1
$\Sigma \lambda_i^{(j)2}$	56	70	2	30	2	6	2

Vì rằng, tổng các bình phương của các “lượng so sánh” ở mỗi so sánh chính là tổng các bình phương của sai lệch ngẫu nhiên của so sánh, cho nên:

- Sai lệch ngẫu nhiên của so sánh là:

$$Q_e(C_1) = \frac{19^2 + 30^2 + 6^2 + 15^2 + 10^2 + 18^2}{56} - \frac{98^2}{6(56)} = 6,17$$

với $6 - 1 = 5$ độ tự do.

.....

- Sai lệch ngẫu nhiên của so sánh 7 là:

$$Q_e(C_7) = \frac{1^2 + 2^2 + (-3)^2 + (-3)^2 + (-1)^2 + 3^2}{2} - \frac{(-1)^2}{6(2)} = 16,42$$

với $6 - 1 = 5$ độ tự do.

Kết quả tính toán các sai lệch ngẫu nhiên của mỗi so sánh ứng với độ tự do của nó được trình bày ở bảng sau đây:

So sánh	1	2	3	4	5	6	7	Tổng
$Q_e(C_j)$	6,17	55,45	25,42	55,47	30,67	12,58	16,42	202,18
Độ tự do	5	5	5	5	5	5	5	35
Biến lượng $S^2(C_j)$	1,234	11,090	5,048	11,094	6,134	2,516	3,284	$S^2 = 5,78$

Để nhận thấy là $\sum Q_e(C_j) = 202,18 = Q_e$ và biến lượng S^2 chính là giá trị trung bình của các biến lượng $S^2(C_j)$.

Kết quả “phân định” cho Q_1 và Q_e cùng với trắc nghiệm F ứng với mỗi so sánh được trình bày ở bảng sau đây:

So sánh	Thí nghiệm								C _j	$\sum n_i \lambda_i^{(j)2}$	$\frac{C_i^2}{\sum n_i \lambda_i^{(j)2}}$	S ² (C _j)	F _j	P _j
	A 44	B 119	c 84	D 51	E 22	F 32	G 33	H 65						
1. A/BCDEFGH	-7	+1	+1	+1	+1	+1	+1	+1	98	6(56)	28,58	1,234	23,16	< 0,005
2. BC/DEFGH	0	+5	+5	-2	-2	-2	-2	-2	609	6(70)	883,05	11,090	79,63	< 0,001
3. B/C	0	+1	-1	0	0	0	0	0	35	6(2)	102,08	5,084	20,08	< 0,01
4. DH/EFG	0	0	0	+3	-2	-2	-2	+3	174	6(30)	168,20	11,094	15,16	< 0,02
5. D/H	0	0	0	+1	0	0	0	-1	-1.4	6(2)	16,33	6,134	2,66	> 0,1
6. E/FG	0	0	0	0	+2	-1	-1	0	-21	6(6)	12,25	2,516	4,87	> 0,05
7. F/G	0	0	0	0	0	+1	-1	0	-1	6(2)	0,08	3,284	0,02	> 0,5
Tổng											1210,57			

Chương 4

TƯƠNG QUAN, HỒI QUY VÀ PHƯƠNG PHÁP THIẾT LẬP CÁC MÔ HÌNH TOÁN HỌC

4.1. SỰ TƯƠNG QUAN VÀ HỒI QUY CỦA HAI ĐẠI LƯỢNG NGẪU NHIÊN

Nếu như hai đại lượng ngẫu nhiên X và Y có mối quan hệ nhân quả $X \leftrightarrow Y$ thì ta nói rằng X và Y có sự tương quan (correlation) nhau. Bây giờ, trong trường hợp X và Y thể hiện sự tương quan rõ rệt thì, hoặc là, chúng ta coi X là nguyên nhân và Y là kết quả thì mối quan hệ $X \leftrightarrow Y$ trở thành $X \rightarrow Y$ và được diễn tả bởi phương trình hồi quy (regression equation) là $y = f(x)$, hoặc là, chúng ta coi Y là nguyên nhân và X là kết quả mối quan hệ $X \leftrightarrow Y$ trở thành $Y \rightarrow X$ và được diễn tả bởi phương trình hồi quy là $x = f(y)$. Phương trình hồi quy chính là mô hình toán học được thiết lập trên cơ sở của các số liệu thực nghiệm thu được. Một sự hồi quy (regression) của các số liệu thực nghiệm trên đồ thị của phương trình hồi quy càng tốt bao nhiêu thì mô hình toán học được thiết lập nên càng gần với bản chất của sự vật bấy nhiêu. Do đó, công việc nghiên cứu về sự tương quan của các đại lượng ngẫu nhiên và các phương án hồi quy của các số liệu thực nghiệm đóng một vai trò hết sức quan trọng trong quá trình nghiên cứu nhằm thiết lập các mô hình toán học mô tả bản chất của sự vật.

4.1.1. Sự tương quan và hồi quy tuyến tính

Nếu như các đại lượng ngẫu nhiên X và Y thể hiện sự hồi quy tuyến tính (linear regression) thì để đánh giá mức độ liên kết của mối quan hệ nhân quả $X \leftrightarrow Y$ chúng ta dùng hệ số tương quan (correlation coefficient, linear correlation coefficient) r (hay r_{xy}) được định nghĩa như sau:

$$r = \frac{K[xy]}{\sqrt{V[x]V[y]}}$$

Ở đây:

- $K[xy]$ là giao biến lượng của X và Y , thường được ký hiệu là K_{xy} , được tính bởi công thức sau đây:

$$K[xy] = K_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{n-1}$$

- $V[x]$ (hay $K[xx]$) là biến lượng của X , thường được ký hiệu là S_x^2 , được tính bởi công thức sau đây:

$$V[x] = S_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}{n-1}$$

- $V[y]$ (hay $K[yy]$) là biến lượng của Y , thường được ký hiệu là S_y^2 , được tính bởi công thức sau đây:

$$V[y] = S_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}}{n-1}$$

Do đó:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

Ngoài ra, với phương pháp phi tham số (non - parametric method) chúng ta cũng có thể dùng hệ số tương quan Spearman hay hệ số tương quan hạng (rank correlation coefficient) để đánh giá mức độ liên kết của mối quan hệ nhân quả $X \leftrightarrow Y$ như sau:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n (x'_i - y'_i)^2}{n(n^2 - 1)}$$

Ở đây: x'_i và y'_i là các hạng được chỉ định từ hạng 1 đến hạng n dựa theo x_i và y_i .

Thực ra, công thức của hệ số tương quan hạng cũng được thiết lập dựa trên căn bản của công thức định nghĩa của hệ số tương quan như sau:

Vì rằng:

$$\bar{x}' = \frac{n+1}{2}; \quad \sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{6}$$

Do đó:

$$\sum_{i=1}^n (x'_i - \bar{x}')^2 = \sum_{i=1}^n \left[K - \frac{n+1}{2} \right]^2 = \sum_{i=1}^n K^2 - \frac{(n+1)^2}{4} = \frac{n(n^2-1)}{12}$$

Tương tự:
$$\sum_{i=1}^n (y'_i - \bar{y}')^2 = \frac{n(n^2-1)}{12}$$

Vì rằng:
$$\sum_{i=1}^n (x'_i - \bar{x}')(y'_i - \bar{y}') = \sum_{i=1}^n x'_i y'_i - \frac{n(n+1)^2}{4} =$$

$$= -\frac{1}{2} \sum_{i=1}^n (x'_i - y'_i)^2 + \sum_{k=1}^n K^2 - \frac{n(n+1)^2}{4} = \frac{n(n^2-1)}{12} - \frac{1}{2} \sum_{i=1}^n (x'_i - y'_i)^2$$

$$\begin{aligned} \text{Do đó: } r_s &= \frac{\sum_{i=1}^n (x'_i - \bar{x}')(y'_i - \bar{y}')}{\sqrt{\sum_{i=1}^n (x'_i - \bar{x}')^2 \cdot \sum_{i=1}^n (y'_i - \bar{y}')^2}} = \frac{\frac{n(n^2-1)}{12} - \frac{1}{2} \sum_{i=1}^n (x'_i - y'_i)^2}{\sqrt{\frac{n(n^2-1)}{12} \cdot \frac{n(n^2-1)}{12}}} = \\ &= 1 - \frac{6 \cdot \sum_{i=1}^n (x'_i - y'_i)^2}{n(n^2-1)} \end{aligned}$$

Chú ý rằng, hệ số tương quan r luôn luôn nằm trong giới hạn là $-1 \leq r \leq 1$ hay $|r| \leq 1$. Trị tuyệt đối $|r|$ càng lớn thì sự thể hiện mức độ liên kết nhân quả $X \leftrightarrow Y$ càng lớn và ngược lại. Điều này còn có nghĩa là, trị tuyệt đối $|r|$ càng lớn thì sự hồi quy của các số liệu thực nghiệm trên đồ thị của phương trình hồi quy càng tốt và ngược lại.

Ví dụ 4.1. Xác định hệ số tương quan với các số liệu (x_i, y_i) đã cho như sau:

x_i	68,8	63,3	75,5	67,2	71,3	72,8	76,5	63,5	69,9	71,4
y_i	167,0	113,3	159,9	153,6	150,8	181,2	173,1	115,4	125,6	166,2

Vì rằng:

$$n = 10; \Sigma x_i = 700; \Sigma x_i^2 = 49208,42;$$

$$y_i = 1506,1; \Sigma y_i^2 = 232139,51; \Sigma x_i y_i = 106215,43$$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 106215,43 - \frac{(700,2)(1506,1)}{10} = 758,308$$

$$\Sigma(x_i - \bar{x})^2 = 49208,42 - \frac{(700,2)^2}{10} = 180,416$$

$$\Sigma(y_i - \bar{y})^2 = 232139,51 - \frac{(1506,1)^2}{10} = 5305,789$$

Do đó:

$$r = \frac{758,308}{\sqrt{180,416 \cdot 5305,789}} = 0,775$$

Để so sánh, chúng ta tiến hành xác định hệ số tương quan hạng r_s như sau :

x_i	68,8	63,3	75,5	67,2	71,3	72,8	76,5	63,5	69,9	71,4
x'_i	4	1	9	3	6	8	10	2	5	7
y_i	167,0	113,3	159,9	153,6	150,8	181,2	173,1	115,4	125,6	166,2
y'_i	8	1	6	5	4	10	9	2	3	7
$x'_i - y'_i$	-4	0	3	-2	2	-2	1	0	2	0

Vì rằng: $\Sigma(x'_i - y'_i)^2 = (-4)^2 + 0^2 + \dots + 0^2 = 42, n = 10$

Do đó: $r_s = 1 - \frac{6(42)}{10(10^2 - 1)} = 0,745$

Để so sánh r và r_s chúng ta dùng trắc nghiệm U như sau (nội dung sẽ được trình bày chi tiết ở mục 4.1.4):

$$U = \frac{\text{Arth}0,775 - \text{Arth}0,745}{\sqrt{\frac{1}{10-3} + \frac{1}{10-3}}} = 0,133; P > 0.5$$

Vậy sự thăng giáng của r và r_s chỉ là một sự thăng giáng ngẫu nhiên, nghĩa là r và r_s là như nhau ($r \sim r_s$).

Chú ý rằng, trong trường hợp hệ số tương quan r tồn tại thì số liệu thực nghiệm sẽ “hồi quy” theo mô hình toán học dạng $y = a + bx$ [hay $y - \bar{y} = b(x - \bar{x})$] nếu coi X là nguyên nhân và Y là kết quả, hoặc $x = a' + b'y$ [hay $x - \bar{x} = b'(y - \bar{y})$] nếu coi Y là nguyên nhân và X là kết quả. Ngoài ra, giữa hệ số tương quan r và hệ số hồi quy (regression coefficient) b và b' có mối quan hệ là $r^2 = b.b'$. Tất cả những vấn đề này sẽ được trình bày chi tiết ở các mục tiếp theo.

4.1.2. Sự tương quan và hồi quy phi tuyến tính

Nếu như các đại lượng ngẫu nhiên X và Y thể hiện sự hồi quy phi tuyến tính (Curvilinear regression) thì nhất thiết phải tiến hành “tuyến tính hóa” bằng cách thay x hoặc y hoặc cả x lẫn y bởi các “đại diện” của chúng như $\sqrt{x}$, $\frac{1}{x}$, $\log x$, e^x , x^2 , $\sqrt{y}$, $\frac{1}{y}$, y^2 , $\ln y$... trước khi tiến hành xác định hệ số tương quan r . Trong trường hợp này, giá trị của hệ số tương quan r thể hiện sự đánh giá mức độ của mối quan hệ nhân quả không phải của “ $X \leftrightarrow Y$ ” mà là của “đại diện của $X \leftrightarrow Y$ ” hoặc “ $X \leftrightarrow$ đại diện của Y ” hoặc “đại diện của $X \leftrightarrow$ đại diện của Y ” tùy theo phép “tuyến tính hóa” mà chúng ta đã thực hiện.

Tương tự như trên, nếu hệ số tương quan r tồn tại thì số liệu thực nghiệm sẽ “hồi quy” theo mô hình toán học như đã trình bày ở trên trong đó x và y được thay thế bởi “đại diện” của x và y dùng để tuyến tính hóa.

Ví dụ 4.2: Xác định hệ số tương quan với các số liệu (x_i, y_i) đã cho sau đây:

x_i	11,3	7,8	16,2	6,3	10,4	7,1	11,7	12,9	8,9	14,6	14,5
y_i	0,261	0,079	2,812	0,029	0,181	0,052	0,425	0,738	0,125	1,882	1,130

Vì rằng:

$$n = 11; \Sigma x_i = 121,7; \Sigma x_i^2 = 1455,15$$

$$\Sigma y_i = 7,714; \Sigma y_i^2 = 13,578; \Sigma x_i y_i = 111,022$$

$$\Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) = 111,022 - \frac{(121,7)(7,714)}{11} = 25,677$$

$$\Sigma(x_i - \bar{x})^2 = 1455,15 - \frac{121,7^2}{11} = 108,705$$

$$\Sigma(y_i - \bar{y})^2 = 13,578 - \frac{7,714^2}{11} = 8,168$$

$$\text{Do đó: } r_{xy} = \frac{25,677}{\sqrt{108,705 \cdot 8,168}} = 0,8617$$

Bây giờ thay vì xét mối quan hệ "x ↔ y" như trên, chúng ta thử xét mối quan hệ "x ↔ lgy" như sau :

x_i	11,3	7,8	16,2	6,3	10,4	7,1	11,7	12,9	8,9	14,6	14,5
$\lg y_i$	-0,583	1,1020	0,449	-1,538	-0,742	-1,284	0,372	-0,132	-0,903	0,275	0,063

Vì rằng :

$$n = 11; \Sigma x_i = 121,7; \Sigma x_i^2 = 1455,15$$

$$\Sigma \lg y_i = -5,879; \Sigma (\lg y_i)^2 = 7,370; \Sigma x_i \lg y_i = -43,741$$

$$\Sigma(x_i - \bar{x})(\lg y_i - \overline{\lg y}) = -43,741 - \frac{121,7(-5,879)}{11} = 21,302$$

$$\Sigma(x_i - \bar{x})^2 = 1455,15 - \frac{121,7^2}{11} = 108,705$$

$$\Sigma(\lg y_i - \overline{\lg y})^2 = 7,370 - \frac{(-5,879)^2}{11} = 4,228$$

$$\text{Do đó: } r_{x \lg y} = \frac{21,302}{\sqrt{108,705 \cdot 4,228}} = 0,9936$$

Để xem r_{xy} và $r_{x \lg y}$ có thực sự khác nhau hay chỉ là một sự thăng giáng ngẫu nhiên chúng ta dùng trắc nghiệm U (nội dung sẽ được trình bày ở mục 4.1.4) như sau:

$$U = \frac{\text{Arth}0,9936 - \text{Arth}0,8617}{\sqrt{\frac{1}{11-3} + \frac{1}{11-3}}} = 3,1416; P < 0,005$$

Vậy sự thăng giáng của r_{xy} và $r_{x \lg y}$ là một sự thăng giáng thực sự, nghĩa là r_{xy} và $r_{x \lg y}$ là khác nhau ($r_{xy} \neq r_{x \lg y}$).

Như vậy rõ ràng là, sự tương quan $x \leftrightarrow \lg y$ thể hiện tốt hơn là sự tương quan $x \leftrightarrow y$. Điều này có nghĩa là, số liệu thực nghiệm thể hiện một sự "hồi quy" phi tuyến tính hợp lý hơn là một sự hồi quy tuyến tính. Trong trường

hợp này, phương trình hồi quy có thể là $lgy = a + bx$ nếu coi x là nguyên nhân và y (hay lgy) là kết quả, hoặc là $x = a' + b'lgy$ nếu coi y (hay lgy) là nguyên nhân và x là kết quả. Mỗi quan hệ giữa b , b' và $r_{x|gy}$ sẽ là $r_{x|gy}^2 = b.b'$. Chú ý rằng trong trường hợp này, để “tuyến tính hóa” phương trình hồi quy chúng ta chỉ cần đặt $u = lgy$, khi đó phương trình hồi quy trở thành $u = a + bx$ hoặc $x = a' + b'u$.

4.1.3. Trường hợp số liệu phân nhóm và mã hóa

Trong nhiều trường hợp, để đơn giản hóa công việc tính toán chúng ta thường tiến hành phân nhóm và mã hóa các số liệu thu được. Trong trường hợp này, hệ số tương quan r được xác định như sau :

$$\text{Giả sử rằng } u = \frac{x - x_0}{\alpha} \text{ và } v = \frac{y - y_0}{\beta}, \text{ tức là } x = x_0 + \alpha u \text{ và } y = y_0 + \beta v$$

Vì rằng:

$$V[x] = V[x_0 + \alpha u] = \alpha^2 V[u]$$

$$V[y] = V[y_0 + \beta v] = \beta^2 V[v]$$

$$K[xy] = K[(x_0 + \alpha u)(y_0 + \beta v)] = K[\alpha u \beta v] = \alpha \beta K[uv]$$

Do đó:

$$r_{xy} = \frac{\alpha \beta K[uv]}{\sqrt{\alpha^2 V[u]} \sqrt{\beta^2 V[v]}} = \frac{K[uv]}{\sqrt{V[u]} \sqrt{V[v]}} = r_{uv}$$

Ví dụ 4.3: Xác định hệ số tương quan với các số liệu (x_i, y_i) đã cho sau đây:

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
8,35	3,50	15,25	12,50	11,25	8,00	11,62	8,50	12,35	5,01	13,15	9,02
8,74	1,49	14,51	9,50	14,50	10,00	11,75	11,00	12,50	7,03	13,25	6,49
9,25	6,40	10,50	6,00	14,23	8,40	12,00	9,00	12,76	7,53	13,26	10,50
9,50	4,50	10,75	2,50	16,25	12,00	13,74	8,51	12,85	6,01	13,40	7,51
9,75	5,00	10,76	5,74	11,35	9,50	16,00	11,50	14,75	12,00	13,50	10,00
14,24	7,00	11,00	8,50	11,50	6,00	12,15	6,00	14,26	10,00	14,00	11,00
13,65	9,50	11,00	5,26	11,50	9,00	12,25	8,05	12,85	9,50	16,00	13,00

Vì rằng: $n = 42$; $\Sigma x_i = 522,23$; $\Sigma x_i^2 = 6652,25$

$\Sigma y_i = 336,41$; $\Sigma y_i^2 = 2987,80$; $\Sigma x_i y_i = 4358,626$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 4358,626 - \frac{(522,23)(336,41)}{42} = 175,688$$

$$\Sigma(x_i - \bar{x})^2 = 6652,25 - \frac{522,23^2}{42} = 158,817$$

$$\Sigma(y_i - \bar{y})^2 = 2987,80 - \frac{336,41^2}{42} = 293,236$$

Do đó: $r_{xy} = \frac{175,688}{\sqrt{158,817 \cdot 293,236}} = 0,814$

Bây giờ chúng ta thử phân nhóm và mã hóa số liệu theo công thức dạng

$$u = \frac{x - 11,5}{1} \text{ và } v = \frac{y - 9}{2} :$$

y	V _i	x										n _j	n _j V _j	n _j V _j ²
		8-9	9-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17				
		8,5	9,5	10,5	11,5	12,5	13,5	14,5	15,5	16,5				
		u _i												
		-3	-2	-1	0	1	2	3	4	5				
0-2} 1	-4	1	0	0	0	0	0	0	0	0	1	-4	16	
2-4} 3	-3	1	0	1	0	0	0	0	0	0	2	-6	18	
4-6} 5	-2	0	2	1	1	1	0	0	0	0	5	-10	20	
6-8} 7	-1	0	1	2	1	4	2	0	0	0	10	-10	10	
8-10} 9	0	0	0	0	5	3	3	2	0	0	13	0	0	
10-12} 11	1	0	0	0	1	0	2	3	0	1	7	7	7	
12-14} 13	2	0	0	0	0	0	0	1	1	2	4	8	16	
n _i		2	3	4	8	8	7	6	1	3	Σ=42	Σ=-15	Σ=87	
n _i u _i		-6	-6	-4	0	8	14	18	4	15	Σ=43			
n _i u _i ²		18	12	4	0	8	28	54	16	75	Σ=215			

$$\text{Vì rằng: } n = 42; \sum n_i u_i = 43; \sum n_i u_i^2 = 215$$

$$\sum n_j V_j = -15; \sum n_j V_j^2 = 87; \sum_i \sum_j n_{ij} u_i v_j = 80$$

$$\sum_i \sum_j n_{ij} (u_i - \bar{u})(v_j - \bar{v}) = 80 - \frac{43(-15)}{42} = 95,357$$

$$\sum n_i (u_i - \bar{u})^2 = 215 - \frac{43^2}{42} = 170,976; \sum n_j (v_j - \bar{v})^2 = 87 - \frac{(-15)^2}{42} = 81,643$$

$$\text{Do đó: } r_{xy} = r_{uv} = \frac{95,357}{\sqrt{170,976 \cdot 81,643}} = 0,807$$

$$\text{Vì rằng: } U = \frac{\text{Arth}0,814 - \text{Arth}0,807}{\sqrt{\frac{1}{42-3} - \frac{1}{42-3}}} = 0,09; P > 0,5$$

Do đó: 0,814 và 0,807 chỉ là các thang giá trị ngẫu nhiên, nghĩa là $0,814 \sim 0,807$.

4.1.4. Thẩm định giá trị của hệ số tương quan

Để thẩm định giá trị của hệ số tương quan của mẫu (r) chúng ta có thể dùng trắc nghiệm t , trắc nghiệm U hay trắc nghiệm F như sau:

a) Trường hợp dùng trắc nghiệm t

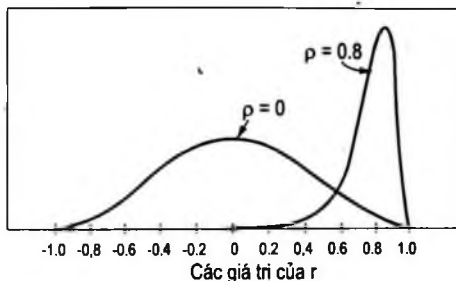
Nếu như hệ số tương quan của quần thể $\rho = 0$ thì đại lượng:

$$\frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \text{ sẽ có phân phối student với } n-2 \text{ độ tự do, nghĩa là:}$$

$$t(n-2) = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Do đó, với giả thiết tương đồng $H_0: \rho = 0$ chúng ta có thể dùng trắc nghiệm t để kiểm tra sự tồn tại (bác bỏ giả thiết H_0) của mối quan hệ nhân quả $X \leftrightarrow Y$. Như thường lệ, nếu giá trị $t(n-2)$ thuộc miền "nội tộc", tức $P > 0,05$, thì giả thuyết $H_0: \rho = 0$ được chấp nhận, vì rằng khi đó $t(n-2) \sim 0$ (giá trị trung bình của phân phối) $\Rightarrow r \sim 0$. Trái lại, nếu giá trị $t(n-2)$ thuộc miền "ngoại tộc", tức $P \leq 0,05$, thì giả thuyết $H_0: \rho = 0$ bị bác bỏ, vì rằng khi đó $t(n-2) \neq 0$ (giá trị trung bình của phân phối) $\Rightarrow r \neq 0$.

Chú ý rằng, trong trường hợp hệ số tương quan của quần thể $\rho \neq 0$ thì sự phân phối của đại lượng r là bất đối xứng (hình 4.1), do đó, trắc nghiệm t là không thích hợp cho những trường hợp mà $\rho \neq 0$. Nói một cách khác, trắc nghiệm t chỉ được dùng để thẩm định mối quan hệ nhân quả $X \leftrightarrow Y$ là tồn tại (bác bỏ giả thuyết $H_0 : \rho = 0$) hay không tồn tại (chấp nhận giả thuyết $H_0 : \rho = 0$). Trong các trường hợp khác, chẳng hạn như kiểm tra giả thuyết $H_0 : \rho = \rho_0$; $H_0 : \rho_1 = \rho_2$ (kể cả trường hợp $\rho_1 = \rho$ và $\rho_2 = 0$, tức $H_0 : \rho = 0$), hoặc xác định khoảng tin tưởng của ρ , thì chúng ta áp dụng phân phối bình thường và dùng trắc nghiệm U để giải quyết vấn đề.



Hình 4.1. Sự phân phối của hệ số tương quan của mẫu (r)

b) Trường hợp dùng trắc nghiệm U

(i) Vì rằng đại lượng $U = \frac{1}{2} \ln \frac{1+r}{1-r} = \text{Arth}r$ có phân phối cận bình thường

dạng $N(\text{Arth}\rho, \frac{1}{\sqrt{n-3}})$, do đó đại lượng $U = \frac{\text{Arth}r - \text{Arth}\rho}{\frac{1}{\sqrt{n-3}}}$ có phân phối

cận bình thường dạng $N(0,1)$. Như vậy:

+ Để thẩm định giá trị của hệ số tương quan r , nghĩa là xét xem mối quan hệ nhân quả $X \leftrightarrow Y$ tồn tại hay không tồn tại, thì chúng ta đặt giả thuyết tương đồng là $H_0 : \rho = 0$ và dùng trắc nghiệm U như sau:

$$U = \frac{\text{Arth}r - \text{Arth}\rho}{\frac{1}{\sqrt{n-3}}} = \frac{\text{Arth}r}{\frac{1}{\sqrt{n-3}}}; P$$

Hoặc có thể dùng khoảng tin tưởng của ρ như sau:

$$-U_p < U < U_p \Rightarrow -U_p < U = \frac{\text{Arth}r - \text{Arth}\rho}{\frac{1}{\sqrt{n-3}}} < U_p$$

Hoặc có thể dùng khoảng tin tưởng của ρ như sau:

$$-U_p < U < U_p \Rightarrow -U_p < U = \frac{\text{Arthr} - \text{Arthp}}{\frac{1}{\sqrt{n-3}}} < U_p$$

$$\Rightarrow \text{Arthr} - \frac{U_p}{\sqrt{n-3}} < \text{Arthp} < \text{Arthr} + \frac{U_p}{\sqrt{n-3}}$$

Nếu khoảng tin tưởng của ρ bao gồm giá trị zero, nghĩa là $\rho \sim 0$, tức mối quan hệ $X \leftrightarrow Y$ không tồn tại (chấp nhận giả thuyết $H_0: \rho = 0$) thường được ký hiệu là $X \leftrightarrow Y$, trái lại, nếu khoảng tin tưởng của ρ không bao gồm giá trị zero, nghĩa là $\rho \neq 0$, tức mối quan hệ $X \leftrightarrow Y$ tồn tại, thường được ký hiệu là $X \leftrightarrow Y$ (bác bỏ giả thuyết $H_0: \rho = 0$).

Trong trường hợp của trắc nghiệm U thì: nếu U thuộc miền "nội tộc", tức $P > 0,05$, thì giả thuyết $H_0: \rho = 0$ được chấp nhận, vì rằng khi đó $U \sim 0$ (giá trị trung bình của phân phối) $\Rightarrow \text{Arthr} \sim 0 \Rightarrow r \sim 0$, trái lại, nếu U thuộc miền "ngoại tộc", tức $P \leq 0,05$, thì giả thuyết $H_0: \rho = 0$ bị bác bỏ, vì rằng khi đó $U \neq 0$ (giá trị trung bình của phân phối) $\Rightarrow \text{Arthr} \neq 0 \Rightarrow r \neq 0$.

+ Để so sánh hệ số tương quan r với hệ số tương quan ρ_0 đã có sẵn thì chúng ta đặt giả thuyết tương đồng là $H_0: \rho = \rho_0$ và dùng trắc nghiệm U hoặc dùng khoảng tin tưởng của ρ tương tự như trên, chú ý là trường hợp của giả thuyết $H_0: \rho = 0$ ở trên chỉ là một trường hợp đặc biệt của giả thuyết $H_0: \rho = \rho_0$ với $\rho_0 = 0$.

$$\text{Trắc nghiệm U: } U = \frac{\text{Arthr} - \text{Arthp}_0}{\frac{1}{\sqrt{n-3}}}; P$$

$$\text{Khoảng tin tưởng của } \rho: \text{Arthr} - \frac{U_p}{\sqrt{n-3}} < \text{Arthp} < \text{Arthr} + \frac{U_p}{\sqrt{n-3}} \quad (\text{xét xem}$$

khoảng tin tưởng của ρ có bao gồm giá trị ρ_0 hay không).

(ii) Vì rằng đại lượng Arthr_1 có phân phối cận bình thường dạng $N(\text{Arthr}_1, \frac{1}{\sqrt{n_1-3}})$ đại lượng Arthr_2 có phân phối cận bình thường dạng $N(\text{Arthr}_2, \frac{1}{\sqrt{n_2-3}})$, do đó, đại lượng:

tương quan r_1 và r_2 với nhau chúng ta đặt giả thuyết tương đồng là $H_0: \rho_1 = \rho_2$ và dùng trắc nghiệm U như sau:

$$U = \frac{(\text{Arthr}_1 - \text{Arthr}_2)}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}; P$$

Nếu U thuộc miền "nội tộc" tức $P > 0,05$, thì giả thuyết $H_0: \rho_1 = \rho_2$ được chấp nhận, vì rằng khi đó $\text{Arthr}_1 - \text{Arthr}_2 \sim 0$ (giá trị trung bình của phân phối) $\Rightarrow \text{Arthr}_1 \sim \text{Arthr}_2 \Rightarrow r_1 \sim r_2$, trái lại, nếu U thuộc miền "ngoại tộc" tức $P \leq 0,05$, thì giả thuyết $H_0: \rho_1 = \rho_2$ bị bác bỏ, vì rằng khi đó $\text{Arthr}_1 - \text{Arthr}_2 \sim 0$ (giá trị trung bình của phân phối) $\Rightarrow \text{Arthr}_1 \neq \text{Arthr}_2 \Rightarrow r_1 \neq r_2$.

Hoặc dùng khoảng tin tưởng của $\text{Arthp}_1 - \text{Arthp}_2$ như sau:

$$-U_p < \frac{(\text{Arthr}_1 - \text{Arthr}_2) - (\text{Arthp}_1 - \text{Arthp}_2)}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} < U_p$$

$$\Rightarrow (\text{Arthr}_1 - \text{Arthr}_2) - U_p \cdot \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} < \text{Arthp}_1 - \text{Arthp}_2 < (\text{Arthp}_1$$

$$\text{Arthp}_2) + U_p \cdot \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}. \text{ Nếu khoảng tin tưởng của } \text{Arthp}_1 - \text{Arthp}_2 \text{ ba}$$

gồm giá trị zero, nghĩa là $\text{Arthp}_1 - \text{Arthp}_2 \sim 0 \Rightarrow \text{Arthp}_1 \sim \text{Arthp}_2 \Rightarrow \rho_1 \sim \rho_2$ (chấp nhận giả thiết $H_0: \rho_1 = \rho_2$), trái lại, nếu khoảng tin tưởng của $\text{Arthp}_1 - \text{Arthp}_2$ không bao gồm giá trị zero, nghĩa là $\text{Arthp}_1 - \text{Arthp}_2 \neq 0 \Rightarrow \text{Arthp}_1 \neq \text{Arthp}_2 \Rightarrow \rho_1 \neq \rho_2$

c) Trường hợp dùng trắc nghiệm F

Vì rằng đại lượng $\frac{\frac{r^2}{1-r^2}}{n-2}$ có phân phối Fisher với 1 và $n-2$ độ tự do, tức

là:

$$F(1, n-2) = \frac{\frac{r^2}{1-r^2}}{n-2}$$

Do đó, chúng ta có thể dùng trắc nghiệm F để thẩm định giá trị của hệ số tương quan r , nghĩa là xét xem mối quan hệ nhân quả $X \leftrightarrow Y$ tồn tại (bác bỏ giả thuyết $H_0: \rho = 0$ nếu $P \leq 0,05$) hay không tồn tại (chấp nhận giả thuyết $H_0: \rho = 0$ nếu $P > 0,05$). Dễ dàng nhận thấy là $t^2(n-2) = F(1, n-2)$. Ngoài ra, qua trắc nghiệm F chúng ta cũng dễ dàng hình dung hơn ý nghĩa của hệ số xác định (coefficient of determination) r^2 là sự thể hiện % mức đóng góp tạo nên sự hồi quy của số liệu thực nghiệm so với % tạo thành từ các thăng giáng ngẫu nhiên của số liệu thực nghiệm được diễn tả bởi hệ số không xác định (coefficient of nondetermination) là $(1 - r^2)$. Điều này sẽ được làm rõ khi xem xét tới vấn đề “sự phù hợp của mô hình toán học”.

Ví dụ 4.4. Thẩm định giá trị của hệ số tương quan ở ví dụ 4.3.

Cách 1: Dùng trắc nghiệm t với giả thuyết $H_0: \rho = 0$.

$$T(42-2) = \frac{0,814}{\sqrt{\frac{1-0,814^2}{42-2}}} = 8,863; P < 0,001$$

Vậy giả thuyết $H_0: \rho = 0$ bị bác bỏ, nghĩa là $\rho \neq 0$, do đó, mối quan hệ nhân quả $X \leftrightarrow Y$ thực sự tồn tại.

Cách 2: Dùng trắc nghiệm U với giả thuyết $H_0: \rho = 0$

$$U = \frac{\text{Arth}0,814}{\frac{1}{\sqrt{42-3}}} = 7,112; P < 0,001 \text{ (kết luận như trên).}$$

Cách 3: Dùng khoảng tin tưởng của ρ với $P = 0,001$

$$\begin{aligned} \text{Arth}0,814 - \frac{3,291}{\sqrt{42-3}} &< \text{Arth}\rho < \text{Arth}0,814 + \frac{3,291}{\sqrt{42-3}} \\ \Rightarrow 1,139 - 0,527 &< \text{Arth}\rho < 1,139 + 0,527 \\ \Rightarrow 0,612 &< \text{Arth}\rho < 1,666 \Rightarrow 0,546 < \rho < 0,931 \end{aligned}$$

Vậy khoảng tin tưởng của ρ là $(0,546; 0,931)$ không bao gồm giá trị zero, nghĩa là $\rho \neq 0$, do đó giả thuyết $H_0: \rho = 0$ bị bác bỏ (kết luận như trên).

Cách 4: Dùng trắc nghiệm F với giả thuyết $H_0: \rho = 0$.

$$F(1; 42 - 4) = \frac{\frac{0,814^2}{1 - 0,814^2}}{\frac{1}{42 - 2}} = 78,552; P < 0,001 \text{ (kết luận như trên).}$$

4.1.5. Xác lập mô hình hồi quy

Nếu kết quả thẩm định giá trị của hệ số tương quan của mẫu (r) cho biết rằng mối quan hệ nhân quả $X \leftrightarrow Y$ tồn tại thì công việc tiếp theo là xác lập mô hình hồi quy của số liệu thực nghiệm. Chú ý rằng hệ số tương quan r được thiết lập trên căn bản của mối quan hệ tuyến tính hoặc đã được “tuyến tính hóa” giữa các yếu tố “nguyên nhân” và “kết quả”, do đó, dạng chung của mô hình hồi quy của hai đại lượng ngẫu nhiên là $y = a + bx$ hay $y - \bar{y} = b(x - \bar{x})$. Ngoài ra, cũng cần để ý là, nếu Y_i là số liệu thực nghiệm và y_i là giá trị được tính từ phương trình hồi quy $y = f(x)$ ứng với giá trị x_i thì ta luôn luôn có được:

$$\sum Y_i = \sum y_i \text{ hay } \sum (Y_i - y_i) = 0 \text{ và } \bar{Y} = \bar{y}$$

Chú ý rằng, đồ thị của phương trình hồi quy $y = f(x)$ luôn luôn đi qua điểm $(\bar{x}, \bar{y})$.

Bây giờ vấn đề đặt ra ở đây là làm thế nào để xác định các hệ số a, b của phương trình hồi quy. Hai phương pháp thường dùng để xác định các hệ số của phương trình hồi quy là phương pháp “bình phương tối thiểu” (The method of least squares) và phương pháp “moment” (the method of moments). Nội dung của hai phương pháp này như sau:

a) Phương pháp bình phương tối thiểu

Với $e_i = Y_i - y_i$ thì $\sum e_i = 0$ và $\sum e_i^2 = \min$, do đó để xác định các hệ số của phương trình hồi quy chúng ta chỉ cần tối thiểu hóa (minimization) hàm $\sum e_i^2$ như sau:

$$\begin{aligned} \sum e_i^2 &= \sum (Y_i - y_i)^2 = \sum [Y_i - (a + bx_i)]^2 = \sum [Y_i^2 + (a + bx_i)^2 - 2Y_i(a + bx_i)] \\ &= \sum Y_i^2 + \sum (a + bx_i)^2 - 2\sum Y_i(a + bx_i) \\ \frac{\partial (\sum e_i^2)}{\partial a} &= \sum 2(a + bx_i) - 2\sum Y_i = 2[\sum (a + bx_i) - \sum Y_i] \\ &= 2[na + b\sum x_i - \sum Y_i] = 0 \Rightarrow na + b\sum x_i = \sum Y_i \end{aligned} \quad (1)$$

$$\begin{aligned}\frac{\partial(\sum e_i^2)}{\partial b} &= \sum 2x_i(a + bx_i) - \sum 2x_i Y_i = 2[\sum x_i(a + bx_i) - \sum x_i Y_i] \\ &= 2[a\sum x_i + b\sum x_i^2 - \sum x_i Y_i] = 0 \Rightarrow a\sum x_i + b\sum x_i^2 = \sum x_i Y_i \quad (2)\end{aligned}$$

Từ (1) và (2) suy ra:

$$\begin{aligned}a &= \frac{\begin{vmatrix} \sum Y_i & \sum x_i \\ \sum x_i Y_i & \sum x_i^2 \end{vmatrix}}{\begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}} = \frac{\sum x_i^2 \cdot \sum Y_i - \sum x_i \cdot \sum x_i Y_i}{n\sum x_i^2 - (\sum x_i)^2} \\ &= \frac{\sum Y_i \left[x_i^2 - \frac{1}{n}(\sum x_i)^2 \right] + \frac{1}{n} \sum Y_i (\sum x_i)^2 - \sum x_i \cdot \sum x_i Y_i}{n \left[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2 \right]} \\ &= \frac{\sum Y_i \left[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2 \right] - \sum x_i \left[\sum x_i Y_i - \frac{1}{n} \sum x_i \cdot \sum Y_i \right]}{n \left[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2 \right]} \\ &= \frac{\sum Y_i}{n} - \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} \cdot \frac{\sum x_i}{n} = \bar{Y} - b\bar{x} \\ b &= \frac{\begin{vmatrix} n & \sum Y_i \\ \sum x_i & \sum x_i Y_i \end{vmatrix}}{\begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}} = \frac{n \cdot \sum x_i Y_i - \sum x_i \cdot \sum Y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{n \left(\sum x_i Y_i - \frac{1}{n} \sum x_i \cdot \sum Y_i \right)}{n \left[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2 \right]} \\ &= \frac{\sum (x_i - \bar{x}) \cdot (Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}\end{aligned}$$

Chú ý rằng, thay vì dùng phương trình hồi quy dạng $y = a + bx$ như trên, chúng ta có thể dùng phương trình hồi quy dạng $y - \bar{Y} = b(x - \bar{x})$ hay $y = \bar{Y} + b(x - \bar{x})$ như sau:

$$\begin{aligned}
\sum e_i^2 &= \sum (Y_i - y_i)^2 = \sum [Y_i - \bar{Y} - b(x_i - \bar{x})]^2 \\
&= \sum [(Y_i - \bar{Y})^2 + b^2(x_i - \bar{x})^2 - 2b(Y_i - \bar{Y})(x_i - \bar{x})] \\
&= \sum (Y_i - \bar{Y})^2 + \sum b^2(x_i - \bar{x})^2 - \sum 2b(x_i - \bar{x})(Y_i - \bar{Y}) \\
\frac{\partial(\sum e_i^2)}{\partial b} &= \sum 2b(x_i - \bar{x})^2 - \sum 2(x_i - \bar{x})(Y_i - \bar{Y}) \\
&= 2 \left[b \sum (x_i - \bar{x})^2 - \sum (x_i - \bar{x})(Y_i - \bar{Y}) \right] = 0 \Rightarrow b = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

Vì rằng đồ thị của phương trình hồi quy $y = a + bx$ đi qua $(\bar{x}, \bar{Y})$, do đó:
 $\bar{Y} = a + b\bar{x} \Rightarrow a = \bar{Y} - b\bar{x}$

b) Phương pháp "moment"

$$\text{Vì rằng: } \sum Y_i = \sum y_i = \sum (a + bx_i) = na + b \sum x_i \quad (1)$$

$$\text{Do đó: } \sum x_i Y_i = \sum x_i y_i = \sum x_i (a + bx_i) = a \sum x_i + b \sum x_i^2 \quad (2)$$

Từ (1) và (2) chúng ta suy ra được a và b giống như trên. Chú ý rằng ở đây chúng ta cũng có thể dùng phương trình hồi quy dạng: $y - \bar{Y} = b(x - \bar{x})$ như sau:

$$\text{Vì rằng: } \sum (Y_i - \bar{Y}) = \sum (y_i - \bar{Y}) = b \sum (x_i - \bar{x})$$

$$\text{Do đó: } \sum (x_i - \bar{x})(Y_i - \bar{Y}) = b \sum (x_i - \bar{x})(x_i - \bar{x}) = b \sum (x_i - \bar{x})^2$$

$$\Rightarrow b = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}; a = \bar{Y} - b\bar{x}$$

Tương tự, các hệ số a' và b' của phương trình hồi quy $x = a' + b'y$ sẽ là:

$$b' = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2}; a' = \bar{x} - b'\bar{Y}$$

$$\text{Rõ ràng là: } b.b' = \frac{[\sum (x_i - \bar{x})(Y_i - \bar{Y})]^2}{\sum (x_i - \bar{x})^2 \cdot \sum (Y_i - \bar{Y})^2} = r^2$$

Trên đây chúng ta xác lập mô hình hồi quy đối với hệ số tương quan r_{xy} .

Tương tự, chúng ta cũng có thể xác lập được các mô hình hồi quy trong trường hợp dùng các “đại diện” để “tuyến tính hóa” mô hình hồi quy, chẳng hạn:

+ Với $r_{x \lg y}$ thì phương trình hồi quy là $\lg y = a + bx$ với các hệ số như sau:

$$b = \frac{\sum (x_i - \bar{x})(\lg y_i - \overline{\lg y})}{\sum (x_i - \bar{x})^2}; \quad a = \overline{\lg y} - b\bar{x}$$

+ Với $r_{\sqrt{x} \cdot y}$ thì phương trình hồi quy là $y = a + b\sqrt{x}$ với các hệ số như sau:

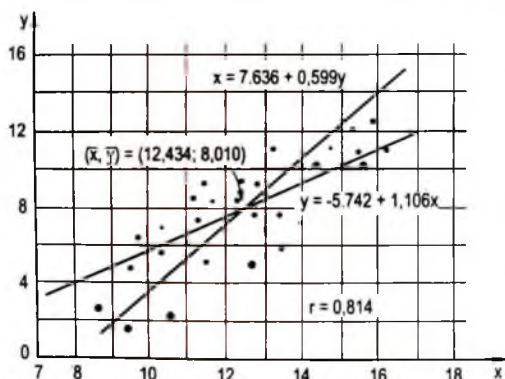
$$b = \frac{\sum (\sqrt{x_i} - \sqrt{\bar{x}})(Y_i - \bar{Y})}{\sum (\sqrt{x_i} - \sqrt{\bar{x}})^2}; \quad a = \bar{Y} - b\sqrt{\bar{x}}$$

+ Với $r_{\frac{1}{x} \cdot y}$ thì phương trình hồi quy là $y = a + b\frac{1}{x}$ với các hệ số như sau:

$$b = \frac{\sum \left[\left(\frac{1}{x_i} \right) - \left(\frac{1}{\bar{x}} \right) \right] (Y_i - \bar{Y})}{\sum \left[\left(\frac{1}{x_i} \right) - \left(\frac{1}{\bar{x}} \right) \right]^2}; \quad a = \bar{Y} - b \left(\frac{1}{\bar{x}} \right)$$

...

Ví dụ 4.5: Xác lập mô hình hồi quy với các số liệu đã cho ở ví dụ 4.3.



Hình 4.2. Đồ thị của các phương trình hồi quy từ số liệu của ví dụ 4.3

Vì rằng $r_{xy} = 0,814$ tồn tại, do đó, mô hình hồi quy $X \rightarrow Y$ có dạng $y = a + bx$ với các hệ số như sau:

$$b = \frac{175,688}{158,817} = 1,106; a = 8,010 - 1,106.12,434 = -5,742$$

Hoặc mô hình hồi quy $Y \rightarrow X$ có dạng $X = a' + b'y$ với các hệ số như sau:

$$b' = \frac{175,688}{293,236} = 0,599; a' = 12,434 - 0,599.8,101 = 7,636$$

Dễ nhận thấy là: $bb' = 1,106.0,599 = 0,662 = r^2$.

Đồ thị của mô hình hồi quy $X \rightarrow Y$ và $Y \rightarrow X$ như sau (hình 4.2).

4.2. SỰ TƯƠNG QUAN VÀ HỒI QUY CỦA NHIỀU ĐẠI LƯỢNG NGẪU NHIÊN

4.2.1. Sự tương quan riêng phần và đa phần

Trên đây chúng ta mới chỉ xem xét mối quan hệ nhân quả của hai đại lượng ngẫu nhiên. Bây giờ chúng ta tiến hành xem xét tới mối quan hệ nhân quả của nhiều (≥ 3) đại lượng ngẫu nhiên. Chú ý rằng, như ở trên đã được đề cập, hệ số tương quan được tính dựa trên căn bản số liệu thực nghiệm hồi quy tuyến tính hay đã được tuyến tính hóa, do đó, trong trường hợp số liệu thực nghiệm thể hiện sự hồi quy phi tuyến tính thì chúng ta phải tiến hành “tuyến tính hóa” trước khi xác định hệ số tương quan.

Trước hết, chúng ta tiến hành xem xét mối quan hệ nhân quả của ba đại lượng ngẫu nhiên X_1 , X_2 và Y như sau:



Bây giờ, nếu như chúng ta muốn đánh giá mức độ của mối quan hệ nhân quả giữa X_1 và X_2 trong điều kiện Y bị cô lập (Y bị cố định hóa) thì chúng ta dùng hệ số tương quan riêng phần (partial correlation coefficient, partial linear correlation coefficient) được định nghĩa như sau:

$$r_{x_1 x_2(y)} = \frac{r_{x_1 x_2} - r_{x_1 y} \cdot r_{x_2 y}}{\sqrt{(1 - r_{x_1 y}^2)(1 - r_{x_2 y}^2)}}$$

Tương tự:

$$r_{x_1 y(x_2)} = \frac{r_{x_1 y} - r_{x_1 x_2} \cdot r_{x_2 y}}{\sqrt{(1 - r_{x_1 x_2}^2)(1 - r_{x_2 y}^2)}}; r_{x_2 y(x_1)} = \frac{r_{x_2 y} - r_{x_2 x_1} \cdot r_{x_1 y}}{\sqrt{(1 - r_{x_2 x_1}^2)(1 - r_{x_1 y}^2)}}$$

Bây giờ, giả sử rằng X_1 , X_2 là các nguyên nhân còn Y là kết quả thì để

đánh giá mức độ của mối quan hệ nhân quả $(X_1, X_2) \rightarrow Y$ thì chúng ta dùng hệ số tương quan đa phần (multiple correlation coefficient, multiple linear correlation coefficient) được định nghĩa như sau:

$$R_{y.12} = \sqrt{1 - (1 - r_{x_1y}^2)(1 - r_{x_2y(x_1)}^2)}$$

Tương tự, hệ số tương quan riêng phần và đa phần của 4 đại lượng ngẫu nhiên X_1, X_2, X_3 và Y được định nghĩa như sau:

$$r_{x_1x_2(x_3y)} = \frac{r_{x_1x_2(y)} - r_{x_1x_3(y)} \cdot r_{x_2x_3(y)}}{\sqrt{(1 - r_{x_1x_3(y)}^2)(1 - r_{x_2x_3(y)}^2)}}$$

...

$$R_{y.123} = \sqrt{1 - (1 - r_{x_1y}^2)(1 - r_{x_2y(x_1)}^2)(1 - r_{x_3y(x_1x_2)}^2)}$$

...

Tương tự, hệ số tương quan riêng phần và đa phần của K đại lượng ngẫu nhiên $X_1, X_2, \dots, X_{K-1}, Y$ được định nghĩa như sau:

$$r_{x_1x_2(x_3 \dots y)} = \frac{r_{x_1x_2(x_4 \dots y)} - r_{x_1x_3(x_4 \dots y)} \cdot r_{x_2x_3(x_4 \dots y)}}{\sqrt{(1 - r_{x_1x_3(x_4 \dots y)}^2)(1 - r_{x_2x_3(x_4 \dots y)}^2)}}$$

...

$$R_{y.12 \dots K-1} = \sqrt{1 - (1 - r_{x_1y}^2)(1 - r_{x_2y(x_1)}^2) \dots (1 - r_{x_{K-1}y(x_1 \dots x_{K-2})}^2)}$$

...

Chú ý rằng, hệ số tương quan đa phần R chính là hệ số tương quan r giữa Y (các số liệu thực nghiệm) và y (các giá trị được tính từ phương trình hồi quy), nghĩa là:

$$R = r_{yY}$$

Thực vậy, để đơn giản hóa vấn đề chúng ta chỉ chứng minh cho trường hợp có hai biến ngẫu nhiên là X và Y như sau:

$$\begin{aligned} r_{yY} &= \frac{\sum (y_i - \bar{y})(Y_i - \bar{Y})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (Y_i - \bar{Y})^2}} = \frac{\sum b(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{b^2 \sum (x_i - \bar{x})^2 \cdot \sum (Y_i - \bar{Y})^2}} = \\ &= \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (Y_i - \bar{Y})^2}} = r_{xy} = R \end{aligned}$$

Chú ý, trong trường hợp chỉ có 2 biến ngẫu nhiên X và Y thì hệ số tương quan r giữa X và Y cũng chính là hệ số tương quan đa phần R kèm theo điều kiện về dấu của hệ số hồi quy b của phương trình hồi quy $y = a + bx$, nghĩa là:

$$r = (\text{dấu của } b)R$$

Ví dụ 4.6: Từ 30 số liệu thực nghiệm (x_{1i}, x_{2i}, y_i) chúng ta tính được các hệ số tương quan của các mối quan hệ nhân quả $X_1 \leftrightarrow X_2$, $X_1 \leftrightarrow Y$, $X_2 \leftrightarrow Y$ như sau:

	X_1	X_2
X_2	0,2094	
Y	-0,7177	-0,4998

Bây giờ, chúng ta tiến hành xác định các hệ số tương quan riêng phần và đa phần như sau:

$$r_{x_1 x_2(y)} = \frac{0,2094 - (-0,7177)(-0,4998)}{\sqrt{(1 - 0,7177^2)(1 - 0,4998^2)}} = -0,2475$$

$$r_{x_1 y(x_2)} = \frac{-0,7177 - (-0,4998)(0,2094)}{\sqrt{(1 - 0,4998^2)(1 - 0,2094^2)}} = -0,7238$$

$$r_{x_2 y(x_1)} = \frac{-0,4998 - (-0,7177)(0,2094)}{\sqrt{(1 - 0,7177^2)(1 - 0,2094^2)}} = -0,5133$$

$$R_{y.12} = \sqrt{1 - (1 - 0,7177^2)(1 - 0,5133^2)} = 0,8018$$

$$R_{x_2.1y} = \sqrt{1 - (1 - 0,4998^2)(1 - 0,2475^2)} = 0,5438$$

$$R_{x_1.2y} = \sqrt{1 - (1 - 0,7177^2)(1 - 0,2475^2)} = 0,7381$$

Như vậy rõ ràng là, sự hồi quy theo mô hình $(X_1, X_2) \rightarrow Y$ là tốt nhất ($R = 0,8018 > 0,7381 > 0,5438$).

Ví dụ 4.7: Nghiên cứu ảnh hưởng của tuổi (X_1), chế độ dinh dưỡng giàu protein (X_2), giàu chất béo (X_3) đối với hàm lượng cholesterol trong máu (Y) ở 54 phụ nữ với các hệ số tương quan tính được như sau:

	X_1	X_2	X_3
X_2	-0,4865		
X_3	-0,5296	0,5784	
Y	0,4737	-0,4249	-0,3135

Bây giờ, nếu như chúng ta muốn đánh giá mức độ của mối quan hệ nhân quả giữa X_1 (tuổi) và Y(cholesterol) trong điều kiện mà sự ảnh hưởng của chế độ dinh dưỡng (X_2 và X_3) bị loại trừ thì chúng ta cần xác định hệ số tương quan riêng phần $r_{x_1y(x_2x_3)}$ như sau:

Vì rằng:

$$r_{x_1y(x_3)} = 0,3820; r_{x_1x_2(x_3)} = -0,2604; r_{x_2y(x_3)} = -0,3145$$

Do đó:

$$r_{x_1y(x_2x_3)} = \frac{0,3820 - (-0,2604)(-0,3145)}{\sqrt{(1-0,2604^2)(1-0,3145^2)}} = 0,3274$$

Ngoài ra, nếu như chúng ta muốn đánh giá mức độ của mối quan hệ nhân quả (X_1, X_2, X_3) $\rightarrow$ Y thì chúng ta cần xác định hệ số tương quan đa phần như sau:

$$\begin{aligned} R_{y.123} &= \sqrt{1 - (1 - r_{x_3y}^2)(1 - r_{x_2y(x_3)}^2)(1 - r_{x_1y(x_2x_3)}^2)} \\ &= \sqrt{1 - (1 - 0,3135^2)(1 - 0,3145^2)(1 - 0,3274^2)} = 0,524 \end{aligned}$$

4.2.2. Thẩm định giá trị của hệ số tương quan riêng phần và đa phần

Giả sử rằng có n số liệu và K biến ngẫu nhiên có thể có các mối quan hệ nhân quả với nhau thì điều chúng ta muốn biết là các mối quan hệ nhân quả nào thì *tồn tại* (bác bỏ giả thuyết tương đồng $H_0: \rho = 0$) và các mối quan hệ nhân quả nào thì *không tồn tại* (chấp nhận giả thuyết tương đồng $H_0: \rho = 0$). Như thường lệ, với $P \leq 0,05$ thì chúng ta bác bỏ giả thuyết tương đồng H_0 , còn $P > 0,05$ thì chúng ta chấp nhận giả thuyết tương đồng H_0 .

a) Thẩm định giá trị của hệ số tương quan riêng phần

Để thẩm định giá trị của hệ số tương quan riêng phần (tồn tại hay không tồn tại), tương tự như việc thẩm định hệ số tương quan, chúng ta có thể dùng trắc nghiệm t, trắc nghiệm F, trắc nghiệm U hay khoảng tin tưởng như sau:

- Dùng trắc nghiệm t với n - K độ tự do:

$$t(n-K) = \frac{r}{\sqrt{\frac{1-r^2}{n-K}}}$$

- Dùng trắc nghiệm F với 1 và n - K độ tự do:

$$F(1, n-K) = \frac{\frac{r^2}{1-r^2}}{\frac{1}{n-K}}$$

- Dùng trắc nghiệm U:

$$U = \frac{\text{Arth}}{\sqrt{\frac{1}{(n-K-1)}}}$$

- Dùng khoảng tin tưởng ở mức xác suất $P \leq 0,05$

$$-U_p < U < U_p \Rightarrow -U_p < \frac{\text{Arth}_r - \text{Arth}_p}{\sqrt{\frac{1}{(n-K-1)}}} < U_p$$

$$\Rightarrow \text{Arth}_r - \frac{U_p}{\sqrt{n-K-1}} < \text{Arth}_p < \text{Arth}_r + \frac{U_p}{\sqrt{n-K-1}}$$

b) Thẩm định giá trị của hệ số tương quan đa phần

Để thẩm định giá trị của hệ số tương quan đa phần tồn tại hay không tồn tại chúng ta dùng trắc nghiệm F với K - 1 và n - K độ tự do như sau :

$$F(K-1, n-K) = \frac{\frac{R^2}{(K-1)}}{\frac{(1-R^2)}{(n-K)}}$$

Chú ý rằng, trắc nghiệm F dùng để thẩm định giá trị của hệ số tương quan đa phần R (hay hệ số tương quan r nếu chỉ có hai biến ngẫu nhiên) chính là trắc nghiệm F dùng để thẩm định mô hình hồi quy tồn tại (phù hợp) hay không tồn tại (không phù hợp), tức thẩm định giá trị của các hệ số hồi quy

(b) là tồn tại (bác bỏ giả thuyết $H_0: \beta = 0$) hay không tồn tại (chấp nhận giả thuyết $H_0: \beta = 0$) mà chúng ta sẽ được thấy ở phần sau :

Ví dụ 4.8: Thẩm định giá trị của hệ số tương quan riêng phần và đa phần ở ví dụ 4.7.

Với $n = 54$ và $K = 4$, việc thẩm định giá trị của hệ số tương quan riêng phần $r_{x_1y(x_2x_3)} = 0,3274$ có thể như sau:

Cách 1: Dùng trắc nghiệm t như sau:

$$t(54 - 4) = \frac{0,3274}{\sqrt{\frac{(1 - 0,3274^2)}{54 - 4}}} = 2,45; P < 0,02$$

Vậy tồn tại mối quan hệ nhân quả $X_1 \xleftrightarrow{(x_2x_3)} Y$

Cách 2: Dùng trắc nghiệm F như sau :

$$F(1, 54 - 4) = \frac{\frac{0,3274^2}{1 - 0,3274^2}}{\frac{1}{54 - 4}} = 6,003; P < 0,02 \text{ (kết luận như trên)}$$

Để nhận thấy là $t^2(50) = 2,452 = 6,003 = F(1,50)$

Cách 3: Dùng trắc nghiệm U như sau:

$$U = \frac{\text{Arth}0,3274}{\sqrt{\frac{1}{54 - 4 - 1}}} = 2,379; P < 0,02 \text{ (kết luận như trên)}$$

Cách 4: Dùng khoảng tin tưởng với $P = 0,02$ như sau :

$$\begin{aligned} \text{Arth}0,3274 - \frac{2,326}{\sqrt{54 - 4 - 1}} &< \text{Arth}\rho < \text{Arth}0,3274 + \frac{2,326}{\sqrt{54 - 4 - 1}} \\ \Rightarrow 0,34 - 0,33 &< \text{Arth}\rho < 0,34 + 0,33 \Rightarrow 0,01 < \text{Arth}\rho < 0,67 \\ \Rightarrow 0,01 < \rho < 0,58 &\Rightarrow \rho \neq 0 \text{ (kết luận như trên).} \end{aligned}$$

Việc thẩm định giá trị của hệ số tương quan đa phần $R_{y,123} = 0,524$ như sau:

$$F(4 - 1, 54 - 4) = \frac{\frac{0,524^2}{4 - 1}}{\frac{1 - 0,524^2}{54 - 4}} = 6,308; P < 0,005$$

Vậy tồn tại mối quan hệ nhân quả $(X_1, X_2, X_3) \rightarrow Y$

4.2.3. Xác lập mô hình hồi quy

Nếu kết quả thẩm định giá trị của hệ số tương quan đa phần $R_{y,12...K-1}$ cho biết rằng mối quan hệ nhân quả $(X_1, X_2, ..., X_{K-1}) \rightarrow Y$ tồn tại thì phương trình hồi quy sẽ là: $y = b_0 + b_1x_1 + b_2x_2 + ... + b_{K-1}x_{K-1}$ hoặc có thể viết dưới dạng sau đây:

$$y - \bar{Y} = b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + ... + b_{K-1}(x_{K-1} - \bar{x}_{K-1})$$

Chú ý rằng, mô hình hồi quy được xác lập dựa trên căn bản của hệ số tương quan đa phần lớn nhất (xem ví dụ 4.6) và thể hiện mối quan hệ tuyến tính hoặc đã được tuyến tính hóa. Để tuyến tính hóa chúng ta dùng các “đại diện” của x_K và y thay vì dùng x_K và y như đã trình bày ở các phần trên.

Để xác định các hệ số của phương trình hồi quy chúng ta có thể dùng phương pháp “bình phương tối thiểu” hoặc phương pháp “moment”. Nội dung căn bản của hai phương pháp này đã được trình bày ở mục 4.1.5. Nội dung thực hành của hai phương pháp này sẽ được trình bày ở mục 4.3.3.

4.3. PHƯƠNG PHÁP THIẾT LẬP CÁC MÔ HÌNH TOÁN HỌC

Xây dựng cho được mô hình toán học (công thức, phương trình, hệ phương trình) về vấn đề mình đang nghiên cứu là hết sức cần thiết và cũng là niềm mong mỏi của các nhà nghiên cứu. Mô hình toán học sẽ giúp khái quát vấn đề ở tầm cao hơn, giúp chúng ta hiểu thấu được bản chất sâu xa hơn của vấn đề nghiên cứu, hỗ trợ đắc lực cho sự tư duy sáng tạo và các dự tính, dự báo khoa học.

Với các phương pháp căn bản trình bày dưới đây, chúng tôi hy vọng các bạn sẽ giảm được ít nhiều khó khăn trong nỗ lực tìm hiểu, xác định tính quy luật của vấn đề đang nghiên cứu, thiết lập được các mô hình toán học cho chúng. Chúng tôi cũng hy vọng rằng các mô hình toán học mà các bạn sẽ xây dựng được là hoàn toàn mới để nó được mãi mãi mang tên các bạn, người đã tìm ra nó.

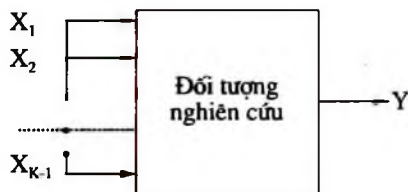
Để có thể thiết lập được một mô hình toán học cần tiến hành các bước sau đây:

- 1 - Mô hình hóa đối tượng nghiên cứu để thu thập số liệu.
- 2 - Dự đoán mô hình toán học trên cơ sở các số liệu thu thập được.
- 3 - Xác định các hệ số của mô hình toán học.
- 4 - Thẩm định sự phù hợp của mô hình toán học.

4.3.1. Mô hình hóa đối tượng nghiên cứu

Bất kỳ kết quả hay hậu quả nào cũng đều có nguyên nhân của nó. Chúng ta dự tính càng đầy đủ số lượng các nguyên nhân gây ra bao nhiêu, đặc biệt là các nguyên nhân chủ đạo, thì chúng ta dự báo càng chính xác kết quả hay hậu quả xảy ra bấy nhiêu. Để thể hiện điều này dưới dạng toán học, chúng ta gọi $X_1, X_2, \dots, X_{K-1}$ là các nguyên nhân tác động gây nên hậu quả hay kết quả Y , thì hàm $y = f(x_1, x_2, \dots, x_{K-1})$ được gọi là mô hình toán học diễn tả mối quan hệ nhân quả $(X_1, X_2, \dots, X_{K-1}) \rightarrow Y$. Chẳng hạn, giả sử như năng suất của cây trồng (Y) bị chi phối bởi 4 yếu tố chủ đạo là nước (X_1), phân (X_2), cần (X_3) và giống (X_4) thì mô hình toán học năng suất của cây trồng sẽ là $y = f(x_1, x_2, x_3, x_4)$.

Để có thể thiết lập được các mô hình toán học thì công việc đầu tiên cần phải tiến hành là mô hình hóa đối tượng nghiên cứu để từ đó cung cấp cho ta các số liệu cần thiết giúp cho công việc xây dựng các mô hình toán học được thuận lợi và chính xác. Việc mô hình hóa đối tượng nghiên cứu có thể được thực hiện bằng cách dựa căn bản vào các mô hình tương tác nhân quả có sẵn trong tự nhiên hoặc là chúng ta phải tự thiết kế, xây dựng nên các mô hình nhằm phục vụ cho công việc nghiên cứu. Sơ đồ tổng quát mô hình hóa đối tượng nghiên cứu như sau (hình 4.3).



Hình 4.3. Mô hình hóa đối tượng nghiên cứu

X: Yếu tố nguyên nhân; Y: Kết quả

4.3.2. Dự đoán các mô hình toán học

Từ các số liệu “vô tri vô giác” thu được làm thế nào có thể dự đoán được các mô hình toán học diễn tả tính quy luật của vấn đề nghiên cứu là một việc làm vô cùng khó khăn. Tuy vậy, với một số phương pháp sẽ được giới thiệu dưới đây, chúng tôi hy vọng rằng các bạn sẽ có được những “hành trang” cần thiết cùng bạn vượt qua những khó khăn ban đầu.

a) Phương pháp “đồ thị thực nghiệm” và “tuyến tính hóa”

Chúng ta có thể dự đoán được các mô hình toán học nhờ đồ thị thực nghiệm được phác họa từ các số liệu thu được. Trong trường hợp này, khó khăn thường gặp phải là việc dự đoán đồ thị của các hàm phi tuyến. Do đó, nếu như đồ thị thực nghiệm có dạng phi tuyến tính thì chúng ta nên tìm cách “tuyến tính hóa” bằng cách thay thế x hoặc y hoặc cả x lẫn y bởi các “đại diện” của chúng như x^2 , x^3 , $\sqrt{x}$, $\frac{1}{x}$, e^x , $\ln x$, ..., $\sqrt{y}$, $\frac{1}{y}$, y^2 , $\ln y$, ..., Chẳng

hạn, nếu như đồ thị thực nghiệm được phác họa bởi các số liệu $(\ln x_i, y_i)$ là tuyến tính thì phương trình hồi quy (mô hình toán học) sẽ là $y = a + b \ln x$, với a, b là các hệ số được xác định bằng phương pháp “bình phương tối thiểu” hay phương pháp “moment”.

Như vậy, nếu như chúng ta dựa vào đồ thị thực nghiệm để dự đoán các mô hình toán học thì cần phác họa số liệu với các trường hợp sau đây trong quá trình tìm kiếm mối quan hệ tuyến tính: (x_i, y_i) , (đại diện của x_i, y_i), $(x_i$, đại diện của y_i) và (đại diện của x_i , đại diện của y_i).

Ví dụ 4.9: Mối quan hệ giữa hiệu điện thế (x) và cường độ dòng điện (y) được đo qua một bóng đèn tungsten 118volt như sau:

X_i, V	2	4	8	16	32	64	128
Y_i, A	0,0245	0,0370	0,0570	0,0855	0,1295	0,2000	0,3035

Rõ ràng là, nếu như chúng ta phác họa đồ thị thực nghiệm bởi (X_i, Y_i) thì đồ thị thực nghiệm sẽ có dạng phi tuyến tính, nhưng nếu chúng ta phác họa đồ thị thực nghiệm bởi $(\lg X_i, \lg Y_i)$ thì đồ thị thực nghiệm sẽ có dạng tuyến tính.

$\lg X_i$	0,30103	0,60206	0,90309	1,20412	1,50515	1,80618	2,10721
$\lg Y_i$	-1,61083	-1,43180	-1,24413	-1,06803	-0,88773	-0,69897	-0,51784

Do đó, mô hình toán học sẽ có dạng là $\lg y = a + b \lg x$

b) Dự đoán mô hình đa thức bậc n

Vì rằng đạo hàm cấp n của đa thức bậc n là một hằng số, nghĩa là $\frac{\partial^n y}{\partial x^n} = c$

(hằng số), do đó, nếu như từ các số liệu thực nghiệm chúng ta xác định được tỷ số

$\frac{\Delta^n Y}{(\Delta x)^n}$ xấp xỉ hằng số thì mô hình toán học sẽ có dạng là

$y = b_0 + b_1x + b_2x^2 + \dots + b_nx^n$. Chú ý rằng, trong trường hợp ∂x là hằng số thì

$\partial^n y = c \cdot \partial x^n$ (hằng số), do đó thay vì xét tỷ số $\frac{\partial^n Y}{\Delta x^n}$ chúng ta chỉ cần xét $\Delta^n Y$:

Nếu Δx là hằng số mà $\Delta^n Y$ cũng là hằng số (hay xấp xỉ hằng số) thì mô hình toán học có dạng đa thức bậc n .

Ví dụ 4.(10): Mối hệ giữa nhiệt độ (x) và môđun xoắn của thép (y) như sau:

X, °C	0	20	40	60	80	100
Y, kg/cm ²	8290	8253	8215	8176	8136	8096

ΔY		-37	-38	-39	-40	-40
$\Delta^2 Y$			-1	-1	-1	0

Vì rằng $\Delta x = 20$ (hằng số) và $\Delta^2 Y$ gần như là hằng số, do đó, mô hình toán học sẽ là đa thức bậc 2 như sau:

$$y = b_0 + b_1x + b_2x^2$$

c) Dự đoán mô hình tỷ lệ thuận, tỷ lệ nghịch

Nếu như tỷ số $\frac{Y_i}{x_i}$ là hằng số (hay xấp xỉ hằng số) thì X và Y có mối

quan hệ tỷ lệ thuận, còn nếu như tích số $x_i \cdot Y_i$ là hằng số (hay xấp xỉ hằng số) thì X và Y có mối quan hệ tỷ lệ nghịch. Bây giờ chúng ta có thể mở rộng vấn đề ra là, nếu như tỷ số (hay tích số) của Y_i hoặc “đại diện của Y” và x_i hoặc “đại diện của x_i ” là hằng số (hay xấp xỉ hằng số) thì Y hoặc “đại diện của Y” và X hoặc “đại diện của X” có mối quan hệ tỷ lệ thuận (hay tỷ lệ nghịch). Chẳng hạn, nếu như $\frac{\ln Y_i}{\sqrt{x_i}} \approx a$ (hằng số) thì mô hình toán học sẽ là

$\ln y = a\sqrt{x}$ hay $y = e^{a\sqrt{x}}$. Trong trường hợp này chúng ta dễ dàng xác định hệ số a , bởi vì như chúng ta đã biết là, đồ thị của phương trình hồi quy luôn luôn đi qua điểm $(\bar{x}, \bar{Y})$.

d) Dự đoán mô hình toán học nhờ sự suy luận hợp lý

Trong nhiều trường hợp, nhờ sự suy luận hợp lý chúng ta cũng có thể dự đoán được các mô hình toán học trong lĩnh vực mà chúng ta đang nghiên cứu và có nhiều kinh nghiệm. Chẳng hạn, mô hình gradient mật độ hay nồng độ có thể dự đoán được là $y = a - bx$, với a là mật độ hay nồng độ ở trung

tâm xuất phát điểm, và x là khoảng cách từ trung tâm đó. Chú ý rằng, ở đây x có thể được thay thế bởi các “đại diện của x ” như $\ln x$, 10^x , e^x , $\sqrt{x}$, x^2 , $x^3 \dots$

e) Dự đoán mô hình “tương tác hình thành các bậc” và “tương tác giữa các yếu tố”

Vì rằng bậc 2 là kết quả tương tác của bậc 1 với nhau, $x.x \rightarrow x^2$, bậc 3 là kết quả tương tác của bậc 1 với bậc 2, $x.x^2 \rightarrow x^3$, bậc 4 là kết quả tương tác của bậc 1 với bậc 3 hoặc của bậc 2 với nhau, $x.x^3$ hay $x^2.x^2 \rightarrow x^4$, ... Do đó, nếu biết vận dụng các khái niệm về “đại diện” và sự “tương tác hình thành bậc” thì chúng ta sẽ có nhiều thuận lợi hơn trong việc dự đoán các mô hình toán học. Chẳng hạn, nếu như đại diện của x là $\sqrt{x}$ với khả năng hình thành bậc 2 thì mô hình toán học được dự đoán sẽ là $y = a + b\sqrt{x} + cx$.

Ngoài các khái niệm về “đại diện” và sự “tương tác hình thành bậc”, trong việc dự đoán các mô hình toán học đa yếu tố (nhiều biến) chúng ta cần biết thêm khái niệm về sự “tương tác giữa các yếu tố”. Chẳng hạn, với hai yếu tố nguyên nhân (X_1 và X_2) thì mô hình toán học đơn giản nhất là $y = b_0 + b_1x_1 + b_2x_2$ nếu không xảy ra tương tác giữa các yếu tố, hoặc là $y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$ nếu xảy ra tương tác giữa các yếu tố. Bây giờ, nếu như yếu tố X_1 hình thành bậc 2 thì mô hình toán học trở thành $y = b_0 + b_1x_1 + b_2x_2 + b_3x_1^2$ (trường hợp không xảy ra tương tác giữa các yếu tố) hoặc là $y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1^2 + b_5x_1^2x_2$ (trường hợp xảy ra tương tác giữa các yếu tố). Bây giờ, nếu như thay x_1 bởi đại diện của nó là $\sqrt{x_1}$ thì mô hình toán học trở thành $y = b_0 + b_1\sqrt{x_1} + b_2x_2 + b_3x_1$ hoặc là $y = b_0 + b_1\sqrt{x_1} + b_2x_2 + b_3x_2\sqrt{x_1} + b_4x_1 + b_5x_1x_2$ (xảy ra tương tác giữa các yếu tố).

Như vậy, trong quá trình tìm kiếm, dự đoán các mô hình toán học chúng ta cần để ý tới các “đại diện”, sự “tương tác hình thành bậc” và sự “tương tác giữa các yếu tố”. Điều này giúp ích cho chúng ta rất nhiều về phương diện lý thuyết cũng như về phương diện thực hành nhằm thiết lập nên các mô hình tốt nhất, phản ánh đúng đắn nhất bản chất của sự vật.

4.3.3. Xác định các hệ số của mô hình toán học

Sau khi dự đoán được mô hình toán học cho sự hồi quy của các số liệu thực nghiệm, công việc tiếp theo là tiến hành xác định các hệ số của mô hình

toán học. Hai phương pháp thường dùng để xác định các hệ số của mô hình toán học là phương pháp “bình phương tối thiểu” và phương pháp “momen”. Cả hai phương pháp này nhằm mục đích là thiết lập “hệ bình thường” (normal equations), giải “hệ bình thường” này chúng ta sẽ có được các hệ số của mô hình toán học. Nội dung căn bản của hai phương pháp này đã được trình bày ở mục 4.1.5. Về phương diện thực hành, chúng ta thường tiến hành thiết lập “hệ bình thường” như sau :

* Đối với mô hình toán học dạng $y = a + bx$

$$a + bx_i = Y_i \rightarrow \left\{ \begin{array}{l} a + bx_i = Y_i \\ ax_i + bx_i^2 = x_i Y_i \end{array} \right\} \rightarrow \left\{ \begin{array}{l} na + b \sum x_i = \sum Y_i \\ a \sum x_i + b \sum x_i^2 = \sum x_i Y_i \end{array} \right.$$

(hệ bình thường).

Vì rằng phương trình hồi quy luôn luôn đi qua điểm $(\bar{x}, \bar{Y})$, do đó, $y = a + bx$ có thể được viết dưới dạng là $y - \bar{Y} = b(x - \bar{x})$. Trong trường hợp này thì:

$$\begin{aligned} b(x_i - \bar{x}) &= Y_i - \bar{Y} \rightarrow b(x_i - \bar{x})^2 = (x_i - \bar{x})(Y_i - \bar{Y}) \rightarrow b \sum (x_i - \bar{x})^2 = \\ &= \sum (x_i - \bar{x})(Y_i - \bar{Y}) \rightarrow b = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}, \text{ và } a = \bar{Y} - b\bar{x} \end{aligned}$$

* Đối với mô hình toán học dạng $y = a + bx + cx^2$.

$$\begin{aligned} a + bx_i + cx_i^2 &= Y_i \rightarrow \left\{ \begin{array}{l} a + bx_i + cx_i^2 = Y_i \\ ax_i + bx_i^2 + cx_i^3 = x_i Y_i \\ ax_i^2 + bx_i^3 + cx_i^4 = x_i^2 Y_i \end{array} \right. \rightarrow \\ &\rightarrow \left\{ \begin{array}{l} na + b \sum x_i + c \sum x_i^2 = \sum Y_i \\ a \sum x_i + b \sum x_i^2 + c \sum x_i^3 = \sum x_i Y_i \\ a \sum x_i^2 + b \sum x_i^3 + c \sum x_i^4 = \sum x_i^2 Y_i \end{array} \right. \end{aligned}$$

(hệ bình thường).

Tương tự như trên, $y = a + bx + cx^2$ có thể được viết dưới dạng là $y - \bar{Y} = b(x - \bar{x}) + c(x^2 - \bar{x}^2)$. Trong trường hợp này thì:

$$\begin{aligned}
 b(x_i - \bar{x}) + c(x_i^2 - \bar{x}^2) &= (Y_i - \bar{Y}) \rightarrow \\
 \rightarrow \begin{cases} b(x_i - \bar{x})^2 + c(x_i - \bar{x})(x_i^2 - \bar{x}^2) = (x_i - \bar{x})(Y_i - \bar{Y}) \\ b(x_i - \bar{x})(x_i^2 - \bar{x}^2) + c(x_i^2 - \bar{x}^2)^2 = (x_i^2 - \bar{x}^2)(Y_i - \bar{Y}) \end{cases} \rightarrow \\
 \rightarrow \begin{cases} b \sum (x_i - \bar{x})^2 + c \sum (x_i - \bar{x})(x_i^2 - \bar{x}^2) = \sum (x_i - \bar{x})(Y_i - \bar{Y}) \\ b \sum (x_i - \bar{x})(x_i^2 - \bar{x}^2) + c \sum (x_i^2 - \bar{x}^2)^2 = \sum (x_i^2 - \bar{x}^2)(Y_i - \bar{Y}) \end{cases}
 \end{aligned}$$

(hệ bình thường), và $a = \bar{Y} - b\bar{x} - c\bar{x}^2$

Chú ý rằng, trong trường hợp này để tránh sự nhầm lẫn đáng tiếc có thể xảy ra trong quá trình tính toán, chẳng hạn, thay vì tính $\bar{x}^2 = \frac{1}{n} \sum x_i^2$, chúng

ta lại tính $\bar{x}^2 = (\frac{1}{n} \sum x_i)^2$, chúng ta nên tiến hành “tuyến tính hóa” mô hình

toán học trước khi thực hiện công việc tính toán. Ở đây, để tuyến tính hóa mô hình toán học $y = a + bx + cx^2$ chúng ta đặt $x = x_1$ và $x^2 = x_2$, bây giờ mô hình toán học sẽ có dạng là $y = a + bx_1 + cx_2$. Như vậy, vấn đề “tuyến tính hóa” các mô hình toán học đóng một vai trò hết sức quan trọng, không những trong công việc dự đoán các mô hình toán học, mà còn trong công việc xác định các hệ số của mô hình toán học cũng như trong công việc thẩm định sự phù hợp của mô hình toán học. Sau đây là một vài ví dụ minh họa cho vấn đề “tuyến tính hóa” của một số mô hình toán học để chúng ta làm quen dần.

$$y = a + \frac{b}{x} \rightarrow y = a + bt \quad \left(\text{đặt } \frac{1}{x} = t \right)$$

$$y = a + be^x \rightarrow y = a + bt \quad \left(\text{đặt } e^x = t \right)$$

$$y = a + b \sin x + c \cos x \rightarrow y = a + bx_1 + cx_2$$

(đặt $\sin x = x_1$, $\cos x = x_2$)

$$y = a + b \ln x \rightarrow y = a + bt \quad \left(\text{đặt } \ln x = t \right)$$

$$y = \frac{1}{a + bx} \rightarrow u = a + bx \quad \left(\text{đặt } \frac{1}{y} = u \right)$$

$$y = e^{a+bx} \rightarrow u = a + bx \quad \left(\text{đặt } \ln y = u \right)$$

$$y = ax^b \rightarrow u = A + bt \quad (\text{đặt } \ln y = u, \ln x = t, \ln a = A)$$

$$y = \frac{x}{a + bx} \rightarrow u = b + at \quad (\text{đặt } \frac{1}{y} = u, \frac{1}{x} = t)$$

$$y = a + bx_1 + cx_2 + dx_1x_2 \rightarrow y = a + bx_1 + cx_2 + dx_3$$

$$(\text{đặt } x_1x_2 = x_3)$$

$$y = a + b \ln x_1 + ce^{x_2} + de^{x_2} \ln x_1 \rightarrow y = a + bt_1 + ct_2 + dt_3$$

$$(\text{đặt } \ln x_1 = t_1, e^{x_2} = t_2, e^{x_2} \ln x_1 = t_3) \dots$$

Tóm lại, có thể nói hầu như tất cả mọi trường hợp của mô hình toán học dạng phi tuyến tính đều có thể “tuyến tính hóa” thành dạng tổng quát sau đây:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_{K-1}x_{K-1}$$

$$\text{hay } y - \bar{Y} = b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + \dots + b_{K-1}(x_{K-1} - \bar{x}_{K-1})$$

$$\text{với } b_0 = \bar{Y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \dots - b_{K-1}\bar{x}_{K-1}$$

Do đó, từ đây trở đi chúng ta chỉ cần tìm hiểu các vấn đề dưới dạng tổng quát trên là đủ.

* Đối với mô hình toán học dạng tổng quát $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_{K-1}x_{K-1}$

$$b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_{K-1}x_{K-1i} = Y_i \rightarrow$$

$$\rightarrow \begin{cases} b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_{K-1}x_{K-1i} = Y_i \\ b_0x_{1i} + b_1x_{1i}^2 + b_2x_{1i}x_{2i} + \dots + b_{K-1}x_{1i}x_{K-1i} = x_{1i}Y_i \\ b_0x_{2i} + b_1x_{1i}x_{2i} + b_2x_{2i}^2 + \dots + b_{K-1}x_{2i}x_{K-1i} = x_{2i}Y_i \\ \dots\dots\dots \\ b_0x_{K-1i} + b_1x_{1i}x_{K-1i} + b_2x_{2i}x_{K-1i} + \dots + b_{K-1}x_{K-1i}^2 = x_{K-1i}Y_i \end{cases}$$

$$\rightarrow \begin{cases} b_0 + b_1\sum x_{1i} + b_2\sum x_{2i} + \dots + b_{K-1}\sum x_{K-1i} = \sum Y_i \\ b_0\sum x_{1i} + b_1\sum x_{1i}^2 + b_2\sum x_{1i}x_{2i} + \dots + b_{K-1}\sum x_{1i}x_{K-1i} = \sum x_{1i}Y_i \\ b_0\sum x_{2i} + b_1\sum x_{1i}x_{2i} + b_2\sum x_{2i}^2 + \dots + b_{K-1}\sum x_{2i}x_{K-1i} = \sum x_{2i}Y_i \\ \dots\dots\dots \\ b_0\sum x_{K-1i} + b_1\sum x_{1i}x_{K-1i} + b_2\sum x_{2i}x_{K-1i} + \dots + b_{K-1}\sum x_{K-1i}^2 = \sum x_{K-1i}Y_i \end{cases}$$

$$\begin{matrix}
 \begin{bmatrix} n & \Sigma x_{1i} & \Sigma x_{2i} & \dots & \Sigma x_{K-li} \\ \Sigma x_{1i} & \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} & \dots & \Sigma x_{1i}x_{K-li} \\ \Sigma x_{2i} & \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 & \dots & \Sigma x_{2i}x_{K-li} \\ \dots & \dots & \dots & \dots & \dots \\ \Sigma x_{K-li} & \Sigma x_{1i}x_{K-li} & \Sigma x_{2i}x_{K-li} & \dots & \Sigma x_{K-li}^2 \end{bmatrix} &
 \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \dots \\ b_{K-1} \end{bmatrix} &
 = &
 \begin{bmatrix} \Sigma Y_i \\ \Sigma x_{1i} Y_i \\ \Sigma x_{2i} Y_i \\ \dots \\ \Sigma x_{K-li} Y_i \end{bmatrix} \\
 \text{A} & \text{B} & & \text{G}
 \end{matrix}$$

Vậy dạng tổng quát của “hệ bình thường” là $AB = G$, với A, B và G là các “ma trận” tương ứng nêu trên.

Trong trường hợp mô hình toán học tổng quát được viết dưới dạng:

$y - \bar{Y} = b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + \dots + b_{K-1}(x_{K-1} - \bar{x}_{K-1})$ thì hệ bình thường được thiết lập như sau:

$$b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2) + \dots + b_{K-1}(x_{K-li} - \bar{x}_{K-1}) = Y_i - \bar{Y}$$

$$\begin{aligned}
 & \left\{ \begin{aligned} & b_1(x_{1i} - \bar{x}_1)^2 + b_2(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) + \dots + b_{K-1}(x_{1i} - \bar{x}_1)(x_{K-li} - \bar{x}_{K-1}) \\ & = (x_{1i} - \bar{x}_1)(Y_i - \bar{Y}) \\ & b_1(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) + b_2(x_{1i} - \bar{x}_1)^2 + \dots + b_{K-1}(x_{2i} - \bar{x}_2)(x_{K-li} - \bar{x}_{K-1}) = \\ \rightarrow & = (x_{2i} - \bar{x}_2)(Y_i - \bar{Y}) \\ & \dots \dots \dots \\ & b_1(x_{1i} - \bar{x}_1)2 + \dots + b_2(x_{2i} - \bar{x}_2)(x_{K-li} - \bar{x}_{K-1}) + \dots + (x_{K-li} - \bar{x}_{K-1})^2 = \\ & = (x_{K-li} - \bar{x}_{K-1})(Y_i - \bar{Y}) \end{aligned} \right.
 \end{aligned}$$

$$\begin{aligned}
 & \left\{ \begin{aligned} & b_1 \sum (x_{1i} - \bar{x}_1)^2 + \dots + b_{K-1} \sum (x_{1i} - \bar{x}_1)(x_{K-li} - \bar{x}_{K-1}) = \\ & = \sum (x_{1i} - \bar{x}_1)(Y_i - \bar{Y}) \\ & b_1 \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) + \dots + b_{K-1} \sum (x_{2i} - \bar{x}_2)(x_{K-li} - \bar{x}_{K-1}) = \\ \rightarrow & = \sum (x_{2i} - \bar{x}_2)(Y_i - \bar{Y}) \\ & \dots \dots \dots \\ & b_1 \sum (x_{K-li} - \bar{x}_{K-1}) + \dots + b_{K-1} \sum (x_{K-li} - \bar{x}_{K-1})^2 = \\ & = \sum (x_{K-li} - \bar{x}_{K-1})(Y_i - \bar{Y}) \end{aligned} \right.
 \end{aligned}$$

$$\begin{bmatrix} \Sigma(x_{li} - \bar{x}_1)^2 & \dots & \Sigma(x_{li} - \bar{x}_1)(x_{K-li} - \bar{x}_{K-1}) \\ \Sigma(x_{li} - \bar{x}_1)(x_{2i} - \bar{x}_2) & \dots & \Sigma(x_{2i} - \bar{x}_2)(x_{K-li} - \bar{x}_{K-1}) \\ \dots & \dots & \dots \\ \Sigma(x_{K-li} - \bar{x}_{K-1}) & \dots & \Sigma(x_{K-li} - \bar{x}_{K-1})^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{K-1} \end{bmatrix} =$$

$$= \begin{bmatrix} \Sigma(x_{li} - \bar{x}_1)(Y_i - \bar{Y}) \\ \Sigma(x_{2i} - \bar{x}_2)(Y_i - \bar{Y}) \\ \vdots \\ \Sigma(x_{K-li} - \bar{x}_{K-1})(Y_i - \bar{Y}) \end{bmatrix}$$

Hệ bình thường $AB = G$ được giải như sau:

$$AB = G \Rightarrow B = A^{-1} G$$

Với A^{-1} là ma trận nghịch đảo của ma trận A . Ma trận nghịch đảo được xác định như sau:

$$[A \quad I] \rightarrow [I \quad A^{-1}]$$

Với I là ma trận đơn vị, chẳng hạn:

$$\begin{bmatrix} a_{11} & a_{12} & \vdots & 1 & 0 \\ a_{21} & a_{22} & \vdots & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & \vdots & \frac{a_{22}}{\Delta} & \frac{-a_{12}}{\Delta} \\ 0 & 1 & \vdots & \frac{-a_{21}}{\Delta} & \frac{a_{11}}{\Delta} \end{bmatrix}$$

$$\Delta = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

4.3.4. Thẩm định và đánh giá mức độ của mô hình toán học

Chúng ta nói rằng mô hình toán học $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_{K-1}x_{K-1}$ hay $y - \bar{Y} = b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + \dots + b_{K-1}(x_{K-1} - \bar{x}_{K-1})$ mà chúng ta đã thiết lập nên là phù hợp với số liệu thực nghiệm thu được, điều này có nghĩa là:

- Giả thuyết tương đồng $H_0: \beta_1 = \beta_2 = \beta_{K-1} = 0$ bị bác bỏ hay giả thuyết tương đồng $H_0: \beta = 0$ bị bác bỏ, do đó có ít nhất là một hệ số hồi quy của mô hình toán học tồn tại, tức $\beta_i \neq 0 \Rightarrow b_i \neq 0$ với $i \neq 0$, hay hệ số tương quan đa phần tồn tại, tức $\rho \neq 0 \Rightarrow R \neq 0$.

- R càng lớn thì mức độ liên kết của mối quan hệ nhân quả: $(X_1, X_2, \dots, X_{K-1}) \rightarrow Y$ càng lớn, khi đó các điểm $(x_{1i}, x_{2i}, x_{K-1i}, Y_i)$ của số liệu thực

nghiệm càng nằm gần với đồ thị của phương trình hồi quy (mô hình toán học), nghĩa là sự phù hợp của mô hình toán học với số liệu thực nghiệm càng cao, và ngược lại.

Trái lại, nếu chúng ta nói rằng mô hình toán học là không phù hợp với số liệu thực nghiệm thu được, thì điều này có nghĩa là giả thuyết tương đồng $H_0 : \beta_1 = \beta_2 = \beta_{K-1} = 0$ được chấp nhận hay giả thuyết tương đồng $H_0 : \rho = 0$ được chấp nhận, do đó, tất cả các hệ số hồi quy của mô hình toán học đều không tồn tại ($b_i \sim 0$ với $i \neq 0$) và tất nhiên hệ số tương quan đa phần cũng không tồn tại ($R \sim 0$).

Trong trường hợp mô hình toán học là phù hợp thì công việc tiếp theo có thể là:

- Kiểm định giá trị của hệ số b_0 là tồn tại hay không tồn tại với giả thuyết tương đồng là $H_0 : \beta_0 = 0$.

- Kiểm định giá trị của hệ số hồi quy b_i , với $i \neq 0$, là tồn tại hay không tồn tại với giả thuyết tương đồng là $H_0 : \beta_i = 0$.

- So sánh các hệ số hồi quy b_i và b_j , với $i \neq 0$ và $j \neq 0$, của mô hình toán học xem là bằng nhau hay khác nhau với giả thuyết tương đồng $H_0 : \beta_i = \beta_j$.

- Xác định khoảng tin tưởng cho giá trị y_0 nào đó của mô hình toán học được tính từ một giá trị $x_0 = (x_1^{(0)}, x_2^{(0)}, \dots, x_{K-1}^{(0)})$.

Trong trường hợp mô hình toán học là không phù hợp hoặc phù hợp nhưng mức độ không cao, tức R không lớn, thì chúng ta nên tiến hành công việc dự đoán một mô hình toán học khác tốt hơn nếu mối quan hệ nhân quả thực sự tồn tại.

a) Trường hợp mô hình toán học dạng $y = b_0 + bx$ (trường hợp $K = 2$) thì "hệ bình thường" là:

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix}$$

$$\begin{aligned} \text{Vì rằng: } \Delta &= \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = n \sum x_i^2 - (\sum x_i)^2 = n \left[\sum x_i^2 - \frac{1}{n} (\sum x_i)^2 \right] \\ &= n \sum (x_i - \bar{x})^2 = nQ_x \quad \text{với } Q_x = \sum (x_i - \bar{x})^2 \end{aligned}$$

Do đó:

$$\begin{bmatrix} b_0 \\ b \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{\sum x_i^2}{\Delta} & \frac{-\sum x_i}{\Delta} \\ \frac{-\sum x_i}{\Delta} & \frac{n}{\Delta} \end{bmatrix} \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix} = \begin{bmatrix} \frac{\sum x_i^2}{nQ_x} & \frac{-\sum x_i}{nQ_x} \\ \frac{-\sum x_i}{nQ_x} & \frac{1}{Q_x} \end{bmatrix} \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix}$$

Nếu mô hình toán học được viết dưới dạng $y - \bar{Y} = b(x - \bar{x})$ thì “hệ bình thường” là:

$$\left[\sum (x_i - \bar{x})^2 \right] [b] = \left[\sum (x_i - \bar{x})(Y_i - \bar{Y}) \right]$$

Suy ra:

$$b = \left[\sum (x_i - \bar{x})^2 \right]^{-1} \left[\sum (x_i - \bar{x})(Y_i - \bar{Y}) \right] =$$

$$= \left[\frac{1}{\sum (x_i - \bar{x})^2} \right] \left[\sum (x_i - \bar{x})(Y_i - \bar{Y}) \right] = \frac{1}{Q_x} \left[\sum (x_i - \bar{x})(Y_i - \bar{Y}) \right]$$

Vì rằng: $e_i = Y_i - y_i$ có phân phối cận bình thường dạng $N(0, \sigma^2)$, do đó, ước lượng tốt nhất của σ^2 sẽ là:

$$S^2 = \frac{\sum (e_i - \bar{e})^2}{n - K} = \frac{\sum (e_i - 0)^2}{n - 2} = \frac{\sum (Y_i - y_i)^2}{n - 2} = \frac{Q_e}{n - 2} \quad \text{với}$$

$$Q_e = \sum (Y_i - y_i)^2$$

Dễ nhận thấy là:

$$\begin{aligned} Y_i - \bar{Y} &= (y_i - \bar{Y}) + (Y_i - y_i) \Rightarrow (Y_i - \bar{Y})^2 = (y_i - \bar{Y})^2 + (Y_i - y_i)^2 + \\ &+ 2(y_i - \bar{Y})(Y_i - y_i) \Rightarrow \sum (Y_i - \bar{Y})^2 = \sum (y_i - \bar{Y})^2 + \sum (Y_i - y_i)^2 \\ &+ 2\sum (y_i - \bar{Y})(Y_i - y_i) \end{aligned}$$

Nhưng vì:

$$\begin{aligned} \sum (y_i - \bar{Y})(Y_i - y_i) &= \sum b(x_i - \bar{x}) [Y_i - \bar{Y} - b(x_i - \bar{x})] = \\ &= \sum b(x_i - \bar{x})(Y_i - \bar{Y}) - \sum b^2(x_i - \bar{x})^2 = b \sum (x_i - \bar{x})(Y_i - \bar{Y}) - b^2 \sum (x_i - \bar{x})^2 \\ &= b.b \sum (x_i - \bar{x})^2 - b^2 \sum (x_i - \bar{x})^2 = 0 \end{aligned}$$

$$\text{Do đó: } \boxed{\sum_{Q_y} (Y_i - \bar{Y})^2 = \sum_{Q_R} (y_i - \bar{Y})^2 + \sum_{Q_e} (Y_i - y_i)^2}$$

Đây là biểu thức căn bản được dùng để thẩm định và đánh giá mức độ phù hợp của mô hình toán học, ở đây:

$$- Q(\text{tổng}) = Q_y = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{1}{n} (\sum Y_i)^2$$

- $Q(\text{hồi quy}) = Q_R = \sum (y_i - \bar{Y})^2$. Cần hết sức chú ý về cách tính Q_R như sau:

+ Nếu $AB = G$ là “hệ bình thường” được thiết lập từ mô hình toán học dạng: $y = b_0 + b_1 x_1 + \dots + b_{K-1} x_{K-1}$ thì:

$$\boxed{Q_R = BG - \frac{1}{n} (\sum Y_i)^2}$$

Do đó ở đây, với mô hình toán học $y = b_0 + bx$ thì:

$$Q_R = [b_0 \quad b] \left[\begin{array}{c} \sum Y_i \\ \sum x_i Y_i \end{array} \right] - \frac{1}{n} (\sum Y_i)^2$$

+ Nếu $AB = G$ là “hệ bình thường” được thiết lập từ mô hình toán học dạng $y - \bar{Y} = b_1(x_1 - \bar{x}) + \dots + b_{K-1}(x_{K-1} - \bar{x}_{K-1})$ thì:

$$\boxed{Q_R = BG}$$

Do đó ở đây, với mô hình toán học $y - \bar{Y} = b(x - \bar{x})$ thì:

$$Q_R = [b] \left[\sum (x_i - \bar{x})(Y_i - \bar{Y}) \right]$$

Dễ nhận thấy là:

$$\begin{aligned} Q_R &= [b_0 \quad b] \left[\begin{array}{c} \sum Y_i \\ \sum x_i Y_i \end{array} \right] - \frac{1}{n} (\sum Y_i)^2 = b_0 \sum Y_i + b \sum x_i Y_i - \frac{1}{n} (\sum Y_i)^2 \\ &= (\bar{Y} - b\bar{x}) \sum Y_i + b \sum x_i Y_i - \frac{1}{n} (\sum Y_i)^2 \\ &= \left(\frac{1}{n} \sum Y_i - b \frac{1}{n} \sum x_i \right) \sum Y_i + b \sum x_i Y_i - \frac{1}{n} (\sum Y_i)^2 \\ &= b \sum x_i Y_i - b \frac{1}{n} \sum x_i \sum Y_i = b \left(\sum x_i Y_i - \frac{1}{n} \sum x_i \sum Y_i \right) \\ &= [b] \left[\sum (x_i - \bar{x})(Y_i - \bar{Y}) \right] \end{aligned}$$

$$- Q (\text{sai lệch ngẫu nhiên}) = Q_e = \sum (Y_i - y_i)^2 = Q_y - Q_R.$$

Chú ý rằng, với $AB = G$ là "hệ bình thường" và A^{-1} là ma trận nghịch đảo của A thì ma trận giao biến lượng sẽ là: $K = S^2 \cdot A^{-1}$

$$\text{Do đó, biến lượng của hệ số } b_i \text{ sẽ là: } S_{b_i}^2 = S^2 C_{ii}$$

$$\text{và giao biến lượng của các hệ số } b_i \text{ và } b_j \text{ sẽ là: } K_{b_i b_j} = S^2 C_{ij}$$

Ở đây, C_{ii} và C_{ij} là các phần tử của ma trận nghịch đảo A^{-1} .

Như vậy biến lượng của các hệ số b_0 và b của mô hình toán học $y = b_0 + bx$ sẽ là:

$$S_{b_0}^2 = S^2 \cdot \frac{\sum x_i^2}{nQ_x}; \quad S_b^2 = S^2 \cdot \frac{1}{Q_x}$$

$$\text{Vì rằng: } \chi^2(n-K) = \frac{(n-K)S^2}{\sigma^2} = \frac{\sum (Y_i - y_i)^2}{\sigma^2} = \frac{Q_e}{\sigma^2}$$

Do đó, với $K = 2$ thì $\chi^2(n-K) = \frac{Q_e}{\sigma^2}$. Chú ý là, đại lượng b có phân phối

cận bình thường dạng $N\left(\beta, \sqrt{\frac{\sigma^2}{Q_x}}\right)$ nên đại lượng $U = \frac{(b-\beta)}{\sqrt{\frac{\sigma^2}{Q_x}}}$ sẽ có phân

phối cận bình thường dạng $N(0, 1)$, do đó: $\chi^2(1) = U^2 = \frac{(b-\beta)^2}{\frac{\sigma^2}{Q_x}}$

Như vậy, tiến trình thẩm định sự phù hợp của mô hình toán học $y = b_0 + bx$ (trường hợp $K = 2$) sẽ như sau:

$$\begin{aligned} F(1, n-2) &= \frac{\frac{\chi^2}{1}}{\frac{\chi^2(n-2)}{(n-2)}} = \frac{(b-\beta)^2}{\frac{\sigma^2}{Q_x}} : \frac{Q_e}{(n-2)\sigma^2} \\ &= \frac{Q_x \cdot (b-\beta)^2}{\sigma^2} \cdot \frac{(n-2)\sigma^2}{Q_e} = \frac{Q_x \cdot (b-\beta)^2}{\frac{Q_e}{n-2}} \end{aligned}$$

Với giả thuyết tương đồng $H_0: \beta = 0$ thì:

$$F(1, n-2) = \frac{Q_x \cdot b^2}{Q_e} = \frac{b \cdot b \sum (x_i - \bar{x})^2}{Q_e} = \frac{[b] \left[\sum x_i - \bar{x} \right] (Y_i - \bar{Y})}{Q_e}$$

Vậy:

$$F(1, n-2) = \frac{Q_R}{Q_e}$$

Tóm lại, để thẩm định sự phù hợp hay tồn tại của mô hình toán học: $y = b_0 + bx$ chúng ta dùng trắc nghiệm F, với $F(1, n-2) = \frac{(n-2)Q_R}{Q_e}$ và giả

thuyết tương đồng là $H_0: \beta = 0$. Như thường lệ, nếu $P > 0,05$ thì giả thuyết tương đồng $H_0: \beta = 0$ được chấp nhận, tức là $b \sim 0$, do đó mô hình toán học $y = b_0 + bx$ không tồn tại hay không phù hợp với số liệu thực nghiệm thu được, trái lại, nếu $P \leq 0,05$ thì giả thuyết tương đồng $H_0: \beta = 0$ bị bác bỏ, tức là $b \neq 0$, do đó, mô hình toán học $y = b_0 + bx$ tồn tại hay phù hợp với số liệu thực nghiệm thu được, với mức độ phù hợp được đánh giá thông qua hệ số tương quan r (ở đây $R = |r|$) như sau:

$$|r| = R = \sqrt{\frac{Q_R}{Q_y}} = \sqrt{1 - \frac{Q_e}{Q_y}}$$

Bảng 4.1. Thẩm định và đánh giá mức độ phù hợp của mô hình toán học $y = b_0 + bx$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Hồi quy	1	Q_R	$S^2 = Q_e / (n - 2)$
Sai lệch ngẫu nhiên	$n - 2$	Q_e	
Tổng	$n - 1$	Q_v	

$$F(1, n - 2) = \frac{\frac{Q_R}{1}}{\frac{Q_E}{n - 2}} ; P$$

$$|r| = \sqrt{\frac{Q_R}{Q_y}} \text{ hay } r = (\text{dấu của } b) \sqrt{\frac{Q_R}{Q_y}}$$

Vì rằng $t^2(n-2) = F(1, n-2)$, do đó trong trường hợp $K = 2$ (trường hợp $y = b_0 + bx$), để thẩm định sự phù hợp của mô hình toán học hay sự tồn tại của hệ số hồi quy b , thay vì dùng trực nghiệm F , chúng ta có thể dùng trực nghiệm t (cũng với giả thuyết tương đồng $H_0: \beta = 0$) như sau:

$$t(n-2) = \frac{b - \beta}{\sqrt{\frac{S^2}{Q_x}}} = \frac{b}{\sqrt{\frac{S^2}{Q_x}}}; P$$

Ngoài ra, chúng ta cũng có thể dùng khoảng tin tưởng của β để kiểm tra giả thuyết $H_0: \beta = 0$ với mức xác suất $P = 0,05$ như sau:

$$b - t_p(n-2) \cdot \sqrt{\frac{S^2}{Q_x}} < \beta < b + t_p(n-2) \cdot \sqrt{\frac{S^2}{Q_x}}$$

Như trên đã nói, trong trường hợp mô hình toán học $y = b_0 + bx$ là phù hợp thì các công việc khác có thể được tiến hành là:

- Thẩm định giá trị của hệ số b_0 với giả thuyết tương đồng $H_0: \beta_0 = 0$ như sau:

$$t(n-2) = \frac{b_0 - \beta_0}{\sqrt{\frac{S^2 \cdot \sum x_i^2}{nQ_x}}} = \frac{b_0}{\sqrt{\frac{S^2 \cdot \sum x_i^2}{nQ_x}}}; P$$

Hoặc có thể dùng khoảng tin tưởng của β_0 với $P = 0,05$ như sau:

$$b_0 - t_p(n-2) \cdot \sqrt{\frac{S^2 \cdot \sum x_i^2}{nQ_x}} < \beta_0 < b_0 + t_p(n-2) \cdot \sqrt{\frac{S^2 \cdot \sum x_i^2}{nQ_x}}$$

- So sánh b_0 và b với giả thuyết tương đồng $H_0: \beta_0 = \beta$ như sau:

Vì rằng: $V[b_0 - b] = V[b_0] + V[b] - 2K[b_0b]$, do đó:

$$S_{b_0-b}^2 = \frac{S^2 \cdot \sum x_i^2}{nQ_x} + \frac{S^2}{Q_x} - 2 \cdot \frac{S^2(-\sum x_i)}{nQ_x} = \frac{S^2(\sum x_i^2 + n + 2\sum x_i)}{nQ_x}$$

Hoặc có thể dùng khoảng tin tưởng của $\beta_0 - \beta$ với $P = 0,05$ như sau:

$$(b_0 - b) - t_p(n-2)S_{b_0-b} < \beta_0 - \beta < (b_0 - b) + t_p(n-2)S_{b_0-b}$$

- Xác định khoảng tin tưởng cho $y_0 = b_0 + bx_0$ như sau:

$$\text{Vì rằng: } V[y_0] = V[b_0 + bx_0] = V[b_0] + x_0^2 V[b] + 2x_0 K[b_0b]$$

Do đó:

$$\begin{aligned}
 S_{y_0}^2 &= \frac{S^2 \sum x_i^2}{nQ_x} + \frac{x_0^2 S^2}{Q_x} + \frac{2x_0 S^2 (-\sum x_i)}{nQ_x} = \frac{S^2}{Q_x} \left(\frac{\sum x_i^2}{n} + x_0^2 - 2x_0 \bar{x} \right) \\
 &= \frac{S^2}{Q_x} \left(\frac{\sum x_i^2}{n} + x_0^2 - 2x_0 \bar{x} + \bar{x}^2 - \bar{x}^2 \right) = \frac{S^2}{Q_x} \left[(x_0 - \bar{x})^2 + \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n} \right)^2 \right] \\
 &= \frac{S^2}{Q_x} \left[(x_0 - \bar{x})^2 + \frac{1}{n} \sum (x_i - \bar{x})^2 \right] = \frac{S^2}{Q_x} \left[(x_0 - \bar{x})^2 + \frac{Q_x}{n} \right] \\
 &= \frac{S^2}{n} + \frac{(x_0 - \bar{x}) S^2}{Q_x} = S^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Q_x} \right]
 \end{aligned}$$

Chú ý rằng, nếu y_0 được tính là $y_0 = \bar{Y} + b(x_0 - \bar{x})$ thì:

$$V[y_0] = V[\bar{Y} + b(x_0 - \bar{x})] = V[\bar{Y}] + (x_0 - \bar{x})^2 V[b]$$

do đó:

$$S_{y_0}^2 = \frac{S^2}{n^2} + \frac{(x_0 - \bar{x})^2 S^2}{Q_x} = S^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Q_x} \right]$$

Vì rằng: $t(n-2) = \frac{y_0 - \mu_0}{S_{y_0}}$. Do đó, khoảng tin tưởng của $\mu_0 = \beta_0 + \beta x_0$ mức

xác suất sai lầm P sẽ như sau:

$$y_0 - t_p(n-2)S_{y_0} < \mu_0 < y_0 + t_p(n-2)S_{y_0} \text{ hay } \mu_0 = y_0 \pm t_p(n-2)S_{y_0}$$

Chú ý:

Hệ bình thường được thiết lập từ mô hình toán học dạng $y - \bar{Y} = b(x - \bar{x})$ sẽ không chứa hệ số b_0 , do đó trong trường hợp này, biến lượng của b_0 được tính như sau:

Vì rằng: $b_0 = Y - b\bar{x}$, do đó: $V[b_0] = V[\bar{Y} - b\bar{x}] = V[\bar{Y}] + \bar{x}^2 V[b]$, từ đó:

$$\begin{aligned}
 S_{b_0}^2 &= \frac{S^2}{n} + \left(\frac{\sum x_i}{n} \right)^2 \cdot \frac{S^2}{Q_x} = \frac{S^2}{nQ_x} \left[Q_x + \frac{(\sum x_i)^2}{n} \right] \\
 &= \frac{S^2}{nQ_x} \left[\sum x_i^2 - \frac{1}{n} (\sum x_i)^2 + \frac{1}{n} (\sum x_i)^2 \right] = \frac{S^2 \cdot \sum x_i^2}{nQ_x}
 \end{aligned}$$

b) Trường hợp mô hình toán học dạng tổng quát $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_{K-1}x_{K-1}$ hay $y - \bar{Y} = b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + \dots + b_{K-1}(x_{K-1} - \bar{x}_{K-1})$.

Chúng minh tương tự như trường hợp $K = 2$, biểu thức căn bản được dùng để thẩm định và đánh giá mức độ phù hợp của mô hình toán học dạng tổng quát vẫn là:

$$\underbrace{\Sigma(Y_i - \bar{Y})^2}_{Q_y} = \underbrace{\Sigma(y_i - \bar{Y})^2}_{Q_R} + \underbrace{\Sigma(Y_i - y_i)^2}_{Q_e}$$

Với giả thuyết tương đồng $H_0: \beta_1 = \beta_2 = \dots = \beta_{K-1} = 0$ thì:

$$F(K-1, n-K) = \frac{\frac{Q_R}{K-1}}{\frac{Q_e}{n-K}}$$

Để đánh giá mức độ phù hợp của mô hình toán học chúng ta dùng hệ số tương quan đa phần R như sau:

$$R = \sqrt{\frac{Q_R}{Q_y}} = \sqrt{1 - \frac{Q_e}{Q_y}}$$

Chú ý rằng, trước đây để thẩm định giá trị của hệ số tương quan đa phần R chúng ta đã dùng trắc nghiệm F với giả thuyết tương đồng $H_0: \rho = 0$ như sau:

$$F(K-1, n-K) = \frac{\frac{R^2}{K-1}}{\frac{1-R^2}{n-K}}$$

Vì rằng:

$$\frac{\frac{R^2}{K-1}}{\frac{1-R^2}{n-K}} = \frac{\frac{Q_R}{Q_y \cdot (K-1)}}{\frac{Q_e}{Q_y \cdot (n-K)}} = \frac{\frac{Q_R}{K-1}}{\frac{Q_e}{n-K}}$$

Do đó trắc nghiệm F dùng để thẩm định giá trị hay sự tồn tại của hệ số tương quan đa phần R cũng chính là trắc nghiệm F được dùng để thẩm định sự phù hợp hay tồn tại của mô hình toán học.

Bảng 4.2. Thẩm định và đánh giá mức độ phù hợp của mô hình toán học $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_{K-1}x_{K-1}$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Hồi quy	K - 1	Q_R	$S^2 = \frac{Q_e}{n - K}$
Sai lệch ngẫu nhiên	n - K	Q_e	
Tổng	n - 1	Q_y	
$F(K-1, n-K) = \frac{\frac{Q_R}{K-1}}{\frac{Q_e}{n-K}}; P$			
$R = \sqrt{\frac{Q_R}{Q_y}}$			

Trước khi trình bày các vấn đề tiếp theo cần ghi nhớ những nội dung sau đây:

+ Về ma trận giao biến lượng: $K = S^2 \cdot A^{-1}$

+ Về biến lượng của b_i : $S_{b_i}^2 = S^2 \cdot C_{ii}$, với C_{ii} là phần tử A^{-1} .

Trong trường hợp “hệ bình thường” không bao gồm b_0 thì biến lượng của b_0 được xác định thông qua đẳng thức $b_0 = \bar{Y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \dots - b_{K-1}\bar{x}_{K-1}$.

+ Về biến lượng của $(b_i - b_j)$: $S_{b_i - b_j}^2 = S^2 \cdot (C_{ii} + C_{jj} - 2C_{ij})$, với C_{ii} , C_{jj} , C_{ij} là các phần tử của A^{-1} .

+ Về biến lượng của $\bar{Y}$: $S_{\bar{Y}}^2 = \frac{S^2}{n}$

+ Về biến lượng của $y_0 = b_0 + b_1x_1^{(0)} + b_2x_2^{(0)} + \dots + b_{K-1}x_{K-1}^{(0)}$ hay:

$$y_0 = \bar{Y} + b_1(x_1^{(0)} - \bar{x}_1) + b_2(x_2^{(0)} - \bar{x}_2) + \dots + b_{K-1}(x_{K-1}^{(0)} - \bar{x}_{K-1})$$

$$S_{y_0}^2 = S^2 \cdot \left[(x_1^{(0)} - \bar{x}_1)^2 C_{11} + (x_2^{(0)} - \bar{x}_2)^2 C_{22} + 2(x_1^{(0)} - \bar{x}_1)(x_2^{(0)} - \bar{x}_2) C_{12} \dots \right]$$

Bây giờ, trong trường hợp mà mô hình toán học phù hợp với số liệu thực nghiệm thu được thì công việc tiếp theo có thể là:

- Xác định giá trị của hệ số b_i với giả thuyết tương đồng $H_0: \beta_i = 0$ như sau:

$$t(n-K) = \frac{b_i - \beta_i}{Sb_i} = \frac{b_i}{Sb_i}; P$$

Hoặc có thể dùng khoảng tin tưởng của β_j với $P \leq 0,05$ như sau :

$$b_i - t_p(n-k)Sb_i < \beta_i < b_i + t_p(n-k)Sb_i$$

- So sánh b_i và b_j với giả thuyết tương đồng $H_0: \beta_i = \beta_j$ như sau:

$$t(n-K) = \frac{(b_i - b_j) - (\beta_i - \beta_j)}{S_{b_i - b_j}} = \frac{b_i - b_j}{S_{b_i - b_j}}; P$$

Hoặc có thể dùng khoảng tin tưởng của $(\beta_i - \beta_j)$ với $P \leq 0,05$ như sau:

$$(b_i - b_j) - t_p(n-k)S_{b_i - b_j} < \beta_i - \beta_j < (b_i - b_j) + t_p(n-2)S_{b_i - b_j}$$

- Xác định khoảng tin tưởng của $\mu_0 = \beta_0 + \beta_1 x_1^{(0)} + \dots + \beta_{K-1} x_{K-1}^{(0)}$ ở mức xác suất sai lầm P như sau:

$$y_0 - t_p(n-2)S_{y_0} < \mu_0 < y_0 + t_p(n-2)S_{y_0} \text{ hay } \mu_0 = y_0 \pm t_p(n-2)S_{y_0}$$

4.3.5. Một số ví dụ minh họa

Ví dụ 4.3.5(1): Thiết lập mô hình toán học của mối quan hệ nhân quả giữa huyết áp và độ tuổi ở phụ nữ theo số liệu nghiên cứu thu được sau đây:

Độ tuổi, x_i	35	45	55	65	75
Huyết áp, Y_i	114	124	143	158	166

+ Mô hình toán học dự đoán: $y = b_0 + bx$ hay $y - \bar{Y} = b(x - \bar{x})$

+ Xác định các hệ số:

$$n = 5, \sum x_i = 275, \sum x_i^2 = 16125, \bar{x} = 55$$

$$\sum Y_i = 705, \sum Y_i^2 = 101341, \sum x_i Y_i = 40155, \bar{Y} = 141$$

$$\text{Kiểm tra: } \sum (x_i + Y_i)^2 = 197776$$

$$\sum x_i^2 + 2\sum x_i Y_i + \sum Y_i^2 = 16125 + 2(40155) + 101341 = 197776$$

$$Q_x = \sum (x_i - \bar{x})^2 = 16125 - \frac{275^2}{5} = 1000$$

$$Q_y = \sum (Y_i - \bar{Y})^2 = 101341 - \frac{705^2}{5} = 1936$$

$$Q_{xy} = \sum (x_i - \bar{x})(Y_i - \bar{Y}) = 40155 - 275 \cdot \frac{705}{5} = 1380$$

Hệ bình thường: $AB = G \Rightarrow B = A^{-1}G$

Với $A = \sum (x_i - \bar{x})^2 = 1000 \Rightarrow A^{-1} = \frac{1}{1000}$

$$B = b$$

$$G = \sum (x_i - \bar{x})(Y_i - \bar{Y}) = 1380$$

Vậy: $b = \frac{1}{1000} \cdot 1380 = 1,38$; $b_0 = \bar{Y} - b\bar{x} = 141 - 1,38(55) = 65,1$

Mô hình toán học đã được thiết lập: $y = 65,1 + 1,38x$

Kiểm tra:

x_i	Y_i	y_i	e_i	e_i^2
35	114	113,4	0,6	0,36
45	124	127,2	3,2	10,24
55	143	141,0	2,0	4,00
65	158	154,8	3,2	10,24
75	166	168,6	2,6	6,76
Tổng			0,0	31,60 = Q_e

+ Thẩm định và đánh giá mức độ phù hợp:

$$Q_R = BG = 1,38(1380) = 1904,4$$

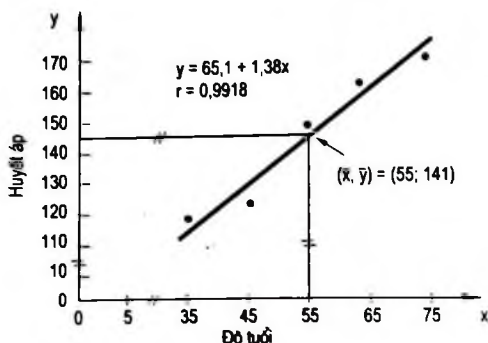
$$Q_e = Q_y - Q_R = 1936 - 1904,4 = 31,6$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Hồi quy	1	1904,4	$S^2 = 10,53$
Sai lệch ngẫu nhiên	3	31,6	
Tổng	4	1936	

$$F(1, 3) = \frac{\frac{1904,4}{1}}{\frac{31,6}{3}} = 180,8; P \ll 0,001$$

$$r = \sqrt{\frac{1904,4}{1936}} = 0,9918$$

- Đồ thị:



+ Thẩm định giá trị của hệ số hồi quy b với giả thuyết $H_0: \beta = 0$ như sau:

$$t(3) = \frac{1,38}{\sqrt{\frac{10,53}{1000}}} = 13,448; P < 0,001$$

Để nhận thấy là $t^2(3) = 13,448^2 = 180,8 = F(1,3)$, do đó trong trường hợp này ($K = 2$) thì không cần phải tiến hành thẩm định giá trị của hệ số hồi quy b.

+ Xác định khoảng tin tưởng của huyết áp cho người phụ nữ có độ tuổi $x_0 = 65$ ở mức xác suất $P = 0,05$.

$$\mu_0 = 154,8 \pm 3,182 \cdot \sqrt{10,53 \left[\frac{1}{5} + \frac{(65-55)^2}{1000} \right]} = 154,8 \pm 5,66$$

Ví dụ 4.3.5(2): Thiết lập mô hình toán học của mối tương quan giữa ngày tuổi và trọng lượng khô (g) của phôi gà theo số liệu nghiên cứu thu được sau đây:

Ngày tuổi, X_i	6	7	8	9	10	11	12	13	14	15	16
Trọng lượng khô Y_i	0,029	0,052	0,079	0,125	0,101	0,261	0,425	0,738	1,130	1,882	2,812

+ Mô hình toán học dự đoán: $\lg y = b_0 + bx$

+ Xác định các hệ số:

X_i	6	7	8	9	10	11	12	13	14	15	16
$\lg Y_i$	-1,538	-1,284	-1,102	-0,903	-0,742	-0,563	0,372	0,132	0,053	0,275	0,449

$$n = 11, \sum x_i = 121, \sum x_i^2 = 1441; \bar{x} = 11$$

$$\sum \lg Y_i = -5,879, \sum (\lg Y_i)^2 = 7,370, \sum x_i (\lg Y_i) = -43,121; \overline{\lg Y} = -0,534$$

Kiểm tra:

$$\begin{aligned} \sum (x_i + \lg Y_i)^2 &= 1362,128 = \sum x_i^2 + 2 \sum x_i (\lg Y_i) + \sum (\lg Y_i)^2 = \\ &= 1441 + 2(-43,121) + 7,370 = 1362,128 \end{aligned}$$

$$Q_x = \sum (x_i - \bar{x})^2 = 1441 - \frac{121^2}{11} = 110$$

$$Q_{\lg y} = \sum (\lg Y_i - \overline{\lg Y})^2 = 7,370 - \frac{(-5,879)^2}{11} = 4,228$$

$$Q_{x \lg y} = \sum (x_i - \bar{x})(\lg Y_i - \overline{\lg Y}) = -43,121 - 121 \cdot \frac{(-5,879)}{11} = 21,548$$

Hệ bình thường: $AB = G \Rightarrow B = A^{-1}G$

$$\text{Với } A = \sum (x_i - \bar{x})^2 = 110 \Rightarrow A^{-1} = \frac{1}{110}$$

$$B = b$$

$$G = \sum (x_i - \bar{x})(\lg Y_i - \overline{\lg Y}) = 21,548$$

$$\text{Vậy: } b = \frac{1}{110} \cdot 21,548 = 0,1959$$

$$b_0 = \overline{\lg Y} - b\bar{x} = -0,534 - 0,1959(11) = -2,689$$

Mô hình toán học đã được thiết lập: $\lg y = -2,689 + 0,1959x$

$$\text{Kiểm tra: } \sum (\lg Y_i - \lg y_i) = 0; \sum (\lg Y_i - \lg y_i)^2 = 0,007 = Q_e$$

Thẩm định và đánh giá mức độ phù hợp:

$$Q_R = BG = 0,1959 \cdot 21,548 = 4,221;$$

$$Q_e = Q_{\lg y} - Q_x = 4,228 - 4,221 = 0,007$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Hồi quy	1	4,221	$S^2 = 0,00078$
Sai lệch ngẫu nhiên	9	0,007	
Tổng	10	4,228	
$F(1, 9) = \frac{\frac{4,221}{1}}{\frac{0,007}{9}} = 5427; P < 0,001$			
$r = \sqrt{\frac{4,221}{4,228}} = 0,9992$			

Chú ý rằng:

$$\begin{aligned} \lg y &= -2,689 + 0,1959x = -\lg 488,65 + (\lg 1,57)x = -\lg 488,65 + \lg 1,57^x = \\ &= \lg \frac{1,57^x}{488,65} = \lg(0,002046)(1,57^x) \end{aligned}$$

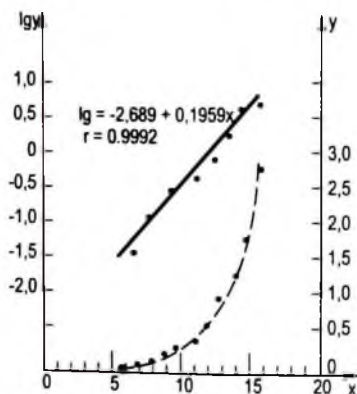
Do đó: $y = (0,002046)(1,57^x)$

Vì rằng: $\ln 1,57 = 0,4511 \Rightarrow 1,57 = e^{0,4511}$

Do đó: $y = 0,002046 \cdot e^{0,4511x}$

Tóm lại, trong trường hợp này mô hình toán học có thể viết dưới ba dạng sau đây : $\lg y = -2,689 + 0,1959x$ hoặc là: $y = (0,002046)(1,57^x)$ hoặc là: $y = 0,002046 \cdot e^{0,4511x}$

- Đồ thị:



+ Thẩm định giá trị của hệ số b_0 với giả thuyết $H_0: \beta_0 = 0$ như sau:

$$t(9) = \frac{-2,689}{\sqrt{\frac{0,00078 \cdot 1441}{11 \cdot 110}}} = 88,23; P < 0,001$$

Xác định khoảng tin tưởng của trọng lượng khô của phôi gà ở $x_0 = 10$ ngày tuổi với mức xác suất $P = 0,05$ như sau:

$$x_0 = 10 \rightarrow \lg y_0 = -2,689 + 0,1959 \cdot 10 = -0,73$$

$$\lg \mu_0 = -0,73 \pm 2,262 \cdot \sqrt{0,00078 \left[\frac{1}{11} + \frac{(10-11)^2}{110} \right]} = -0,73 \pm 0,02$$

hay $-0,75 < \lg \mu_0 < -0,71 \Rightarrow -\lg 0,178 < \lg \mu_0 < \lg 0,195 \Rightarrow 0,178 < \mu_0 < 0,195$

Ví dụ 4.3.5(3): Thiết lập mô hình toán học từ các số liệu (X_i, Y_i) sau đây:

X_i	-3	-2	-1	0	1	2	3
Y_i	-10	0	4	5	4	2	-2

+ Mô hình toán học dự đoán: $y = b_0 + b_1x + b_2x^2$

+ Xác định các hệ số:

Hệ số bình thường:

$$\begin{matrix} \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} & \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} & = & \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum x_i^2 Y_i \end{bmatrix} \\ \text{A} & \text{B} & \text{C} \end{matrix}$$

$$n = 7; \sum x_i = 0; \sum x_i^2 = 28; \sum x_i^3 = 0; \sum x_i^4 = 196$$

$$\sum Y_i = 3; \sum x_i Y_i = 28; \sum x_i^2 Y_i = -92$$

$$\begin{matrix} \begin{bmatrix} 7 & 0 & 28 \\ 0 & 28 & 0 \\ 28 & 0 & 196 \end{bmatrix} & \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} & = & \begin{bmatrix} 3 \\ 28 \\ -92 \end{bmatrix} & \Rightarrow & \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} & = & \begin{bmatrix} 5,38 \\ 1 \\ -1,24 \end{bmatrix} \\ \text{A} & \text{B} & \text{G} \end{matrix}$$

Mô hình toán học đã được thiết lập: $y = 5,38 + x - 1,24x^2$

+ Thẩm định và đánh giá mức độ phù hợp:

$$\sum Y_i^2 = 165; Q_y = \sum (Y_i - \bar{Y})^2 = 165 - \frac{3^2}{7} = 163,71$$

$$BG = (5,38)(3) + (1)(28) + (-1,24)(-92) = 158,22$$

$$Q_R = BG - \frac{(\sum Y_i)^2}{n} = 158,22 - \frac{165^2}{7} = 156,93$$

$$Q_c = Q_y - Q_R = 163,71 - 156,93 = 6,78$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Hồi quy	2	156,93	$S^2 = 1,695$
Sai lệch ngẫu nhiên	4	6,78	
Tổng	6	163,71	

$$F(2, 4) = \frac{\frac{156,93}{2}}{\frac{6,78}{4}} = 46,29 ; P < 0,005$$

$$R = \sqrt{\frac{156,93}{163,71}} = 0,9791$$

+ Thẩm định giá trị của các hệ số của mô hình toán học bằng cách dùng khoảng tin tưởng ở mức xác suất sai lầm $P = 0,05$ như sau:

Ma trận giao biến lượng: $K = S^2 \cdot A^{-1}$

$$K = 1,695 \cdot \begin{bmatrix} 0,333 & 0 & -0,048 \\ 0 & 0,036 & 0 \\ -0,048 & 0 & 0,012 \end{bmatrix} = \begin{bmatrix} 0,565 & 0 & -0,081 \\ 0 & 0,061 & 0 \\ -0,081 & 0 & 0,02 \end{bmatrix}$$

$$\beta_0 = 5,38 \pm 2,776 \cdot \sqrt{0,565} \rightarrow \beta_0 \in (3,29; 7,47)$$

$$\beta_1 = 1 \pm 2,776 \cdot \sqrt{0,061} \rightarrow \beta_1 \in (0,31; 1,69)$$

$$\beta_2 = -1,24 \pm 2,776 \cdot \sqrt{0,02} \rightarrow \beta_2 \in (-1,64; -0,84)$$

+ Xác định khoảng tin tưởng của μ_0 ứng với $x_0 = 2$ ở mức xác suất sai lầm $P = 0,05$ như sau:

$$x_0 = 2 \rightarrow y_0 = 5,38 + 2 - 1,24 \cdot 2^2 = 2,42$$

$$\text{Vì rằng: } V[y_0] = V[b_0 + b_1x_0 + b_2x_0^2] = V[b_0] + x_0^2V[b_1] + x_0^4V[b_2] + \\ + 2x_0K[b_0b_1] + 2x_0^2K[b_0b_2] + 2x_0^3K[b_1b_2]$$

Do đó:

$$S_{y_0}^2 = 0,565 + 2^2 \cdot 0,061 + 2^4 \cdot 0,02 + 2 \cdot 2 \cdot 0 + 2 \cdot 2^3(-0,081) + 2 \cdot 2^3 \cdot 0 = 0481$$

$$\mu_0 = 2,42 \pm 2,776 \cdot \sqrt{0,481} = 2,42 \pm 1,93$$

Ví dụ 4.3.5(4): Thiết lập mô hình toán học của mối quan hệ nhân quả giữa hàm lượng lân vô cơ (X_1), lân hữu cơ (X_2) và lân dễ hấp thu cho thực vật (Y) trong 18 mẫu đất (ppm) như sau:

X_{1i}	0,4	0,4	3,1	0,6	4,7	1,7	9,4	10,1	11,6	12,6	10,9	23,1	23,1
X_{2i}	53	23	19	34	24	65	44	31	29	58	37	46	50
Y_i	64	60	71	61	54	77	81	93	93	51	76	96	77

X_{1i}	21,6	23,1	1,9	26,8	29,9
X_{2i}	44	56	36	58	51
Y_i	93	95	54	168	99

+ Mô hình toán học dự đoán: $y = b_0 + b_1x_1 + b_2x_2$

$$\text{Hay: } y - \bar{Y} = b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2)$$

+ Xác định các hệ số:

Hệ bình thường:

$$\begin{bmatrix} \sum (x_{1i} - \bar{x}_1)^2 & \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \\ \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) & \sum (x_{2i} - \bar{x}_2)^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum (x_{1i} - \bar{x}_1)(Y_i - \bar{Y}) \\ \sum (x_{2i} - \bar{x}_2)(Y_i - \bar{Y}) \end{bmatrix}$$

$$n = 18; \sum x_{1i} = 215; \sum x_{2i} = 758; \sum Y_i = 1463$$

$$\sum x_{1i}^2 = 4321,02; \sum x_{2i}^2 = 35076,00; \sum Y_i^2 = 131299,00$$

$$\sum x_{1i}x_{2i} = 10139,50; \sum x_{1i}Y_i = 20706,20; \sum x_{2i}Y_i = 63825,00$$

$$\bar{x}_1 = 11,94; \bar{x}_2 = 42,11; \bar{Y} = 81,28$$

$$Q_{x_1} = \sum (x_{1i} - \bar{x}_1)^2 = 4321,02 - \frac{215^2}{18} = 1752,96$$

$$Q_{x_2} = \sum (x_{2i} - \bar{x}_2)^2 = 35076,00 - \frac{758^2}{18} = 3155,78$$

$$Q_y = \sum (Y_i - \bar{Y})^2 = 131299,00 - \frac{1463^2}{18} = 12389,61$$

$$Q_{x_1x_2} = \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = 10139,50 - 215 \cdot \frac{758}{18} = 1085,61$$

$$Q_{x_1y} = \sum (x_{1i} - \bar{x}_1)(Y_i - \bar{Y}) = 20706,20 - 215 \cdot \frac{1463}{18} = 3231,48$$

$$Q_{x_2y} = \sum (x_{2i} - \bar{x}_2)(Y_i - \bar{Y}) = 63825,00 - 758 \cdot \frac{1463}{18} = 2216,44$$

$$\begin{matrix} \begin{bmatrix} 1752,96 & 1085,61 \\ 1085,61 & 3155,78 \end{bmatrix} & \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} & = & \begin{bmatrix} 3231,48 \\ 2216,44 \end{bmatrix} & \Rightarrow & \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} & = & \begin{bmatrix} 1,7898 \\ 0,0866 \end{bmatrix} \\ A & B & & G & & & & \end{matrix}$$

$$b_0 = \bar{Y} - b_1\bar{x}_1 + b_2\bar{x}_2 = 81,28 - 1,7898 \cdot 11,94 - 0,0866 \cdot 42,11 = 56,26$$

Vậy mô hình toán học đã được thiết lập: $y = 56,28 + 1,7898x_1 + 0,0866x_2$

$$\text{Kiểm tra: } \sum (Y_i - y_i) = 0; \sum (Y_i - y_i)^2 = 6414 (= Q_e)$$

+ Thẩm định và đánh giá mức độ phù hợp:

$$Q_R = BG = [1,7898; 0,0866] \cdot \begin{bmatrix} 3231,48 \\ 2216,44 \end{bmatrix} = 5975,65$$

$$Q_e = Q_y - Q_R = 12389,61 - 5975,65 = 6413,96$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Hồi quy	2	5975,65	$S^2 = 427,6$
Sai lệch ngẫu nhiên	15	6413,96	
Tổng	17	12389,61	
$F(2, 15) = \frac{\frac{5975,65}{2}}{\frac{6413,96}{15}} = 6,99; P < 0,01$			
$R = \sqrt{\frac{5975,65}{12389,61}} = 0,6945$			

+ Thẩm định giá trị của các hệ số hồi quy b_1 và b_2 bằng cách dùng khoảng tin tưởng ở mức xác suất sai lầm $P = 0,05$ như sau:

Ma trận giao biến lượng: $K = S^2 \cdot A^{-1}$

$$K = 427,6 \cdot \begin{bmatrix} 0,0007249 & -0,0002494 \\ -0,0002494 & 0,0004027 \end{bmatrix} = \begin{bmatrix} 0,3100 & -0,1066 \\ -0,1066 & 0,1722 \end{bmatrix}$$

$$\beta_1 = 17898 \pm 2,131 \cdot \sqrt{0,3100} = 17898 \pm 1,1865 \rightarrow \beta_1 \in (0,6033; 2,9763)$$

$$\beta_2 = 0,0866 \pm 2,131 \cdot \sqrt{0,1722} = 0,0866 \pm 0,8843 \rightarrow \beta_2 \in (-0,7977; 0,9709)$$

Vậy $\beta_1 \neq 0$, còn $\beta_2 \sim 0$, do đó:

- Hoặc là, mối quan hệ nhân quả $X_2 \rightarrow Y$ không tồn tại, do đó có thể loại bỏ yếu tố X_2 ra khỏi mô hình nghiên cứu.

- Hoặc là, tồn tại mối quan hệ nhân quả kiểu “đại diện của X_2 ” $\rightarrow Y$, do đó, cần tiến hành dự đoán và xác định “đại diện của X_2 ”.

+ Xác định khoảng tin tưởng của μ_0 ứng với $x_1^{(0)} = 4,7$ và $x_2^{(0)} = 24$ ở mức xác suất sai lầm $P = 0,05$ như sau:

$$x_0 = (4,7; 24) \rightarrow y_0 = 56,28 + 1,7898 \cdot 4,7 + 0,0866 \cdot 24 = 66,77.$$

Vì rằng:

$$\begin{aligned} V[y_0] &= V[\bar{Y} + b_1(x_1^{(0)} - \bar{x}_1) + b_2(x_2^{(0)} - \bar{x}_2)] \\ &= V[\bar{Y}] + (x_1^{(0)} - \bar{x}_1)^2 V[b_1] + (x_2^{(0)} - \bar{x}_2)^2 V[b_2] + \\ &\quad + 2(x_1^{(0)} - \bar{x}_1)(x_2^{(0)} - \bar{x}_2)K[b_1, b_2] \end{aligned}$$

Do đó:

$$\begin{aligned} S_{y_0}^2 &= \frac{427,6}{18} + (4,7 - 11,94)^2 \cdot 0,31 + (24 - 42,11)^2 \cdot 0,1722 + \\ &\quad + 2(4,7 - 11,94)(24 - 42,11)(-0,1066) = 68,5278 \\ \mu_0 &= 66,77 \pm 2,131 \sqrt{68,5278} = 66,77 \pm 17,64. \end{aligned}$$

Ví dụ 4.3.5(5): Thiết lập mô hình toán học của mối quan hệ nhân quả giữa hàm lượng lân vô cơ (X_1), lân hữu cơ tan trong dung dịch K_2CO_3 và bị thủy phân bởi hypobromite (X_2), lân hữu cơ tan trong dung dịch K_2CO_3 và không bị thủy phân bởi hypobromite (X_3) và lân dễ hấp thu cho thực vật (Y) trong 18 mẫu đất (ppm) như sau:

X_{1i}	X_{2i}	X_{3i}	Y_i	X_{1i}	X_{2i}	X_{3i}	Y_i
0,4	53	158	64	10,9	37	111	76
0,4	23	163	60	23,1	46	114	96
3,1	19	37	71	23,1	50	134	77
0,6	34	157	61	21,6	44	73	93
4,7	24	59	54	23,1	56	168	95
1,7	65	123	77	1,9	36	143	54
9,4	44	46	81	26,8	58	202	168
10,1	31	117	93	29,9	51	124	99
11,6	29	173	93				
12,6	58	112	51				

+ Mô hình toán học dự đoán: $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$

Hay: $y - \bar{Y} = b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + b_3(x_3 - \bar{x}_3)$

+ Xác định các hệ số:

$$\begin{matrix} \begin{bmatrix} Q_{x_1} & Q_{x_1x_2} & Q_{x_1x_2x_3} \\ Q_{x_1x_2} & Q_{x_2} & Q_{x_2x_3} \\ Q_{x_1x_3} & Q_{x_2x_3} & Q_{x_3} \end{bmatrix} & \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} & = & \begin{bmatrix} Q_{x_1y} \\ Q_{x_2y} \\ Q_{x_3y} \end{bmatrix} \\ A & B & G \end{matrix}$$

Tính toán tương tự như ở ví dụ 4.3.5(4) sẽ được:

$$Q_{x_3} = 35572; Q_{x_1x_3} = 1200; Q_{x_2x_3} = 3364; Q_{x_3y} = 7593$$

Do đó:

$$\begin{matrix} \begin{bmatrix} 1752,96 & 1085,61 & 1200,00 \\ 1085,61 & 3155,78 & 3364,00 \\ 1200,00 & 3364,00 & 35572,00 \end{bmatrix} & \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} & = & \begin{bmatrix} 3231,48 \\ 2216,44 \\ 7593,00 \end{bmatrix} & \Rightarrow & \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} & = & \begin{bmatrix} 1,7848 \\ -0,0834 \\ 0,1611 \end{bmatrix} \\ A & B & G \end{matrix}$$

$$b_0 = \bar{Y} - b_1\bar{x}_1 - b_2\bar{x}_2 - b_3\bar{x}_3 = 81,28 - 1,7848(11,94) - (-0,0834)(42,11) - 0,1611(123) = 43,67$$

Mô hình toán học đã được thiết lập:

$$y = 43,67 + 1,7848x_1 - 0,0834x_2 + 0,1611x_3$$

+ Thẩm định và đánh giá mức độ phù hợp:

$$Q_R = BG = [1,7848; -0,0834; 0,1611] \begin{bmatrix} 3231,48 \\ 2216,44 \\ 7593,00 \end{bmatrix} = 6805,93$$

$$Q_e = Q_y - Q_R = 12389,61 - 6805,93 = 5583,68$$

Nguồn biến động	Độ tự do	Tổng các bình phương	Biến lượng
Hồi quy	3	6805,93	$S^2 = 398,83$
Sai lệch ngẫu nhiên	14	5583,68	
Tổng	17	12389,61	
$F(3, 14) = \frac{\frac{6805,93}{3}}{\frac{5583,68}{14}} = 5,69; P < 0,01$			
$R = \sqrt{\frac{6805,93}{12389,61}} = 0,7412$			

Thẩm định giá trị của các hệ số hồi quy b_1 , b_2 và b_3 bằng cách dùng khoảng tin tưởng ở mức xác suất sai lầm $P = 0,05$ như sau:

Ma trận giao biến lượng: $K = S^2 \cdot A^{-1}$

$$K = 398,83 \cdot \begin{bmatrix} 0,0007249 & -0,0002483 & -0,0000010 \\ -0,0002483 & 0,0004375 & -0,0000330 \\ -0,0000010 & -0,0000330 & 0,0000313 \end{bmatrix}$$

$$= \begin{bmatrix} 0,2891 & -0,0990 & -0,004 \\ -0,0990 & 0,1745 & -0,0132 \\ -0,0004 & -0,0132 & 0,0125 \end{bmatrix}$$

$$\beta_1 = 17898 \pm 2,145 \cdot \sqrt{0,2891} = 1,7848 \pm 1,1533 \rightarrow \beta_1 \in (0,6315; 2,9381)$$

$$\beta_2 = -0,0834 \pm 2,145 \cdot \sqrt{0,1745} = -0,0834 \pm 0,8960 \rightarrow \beta_2 \in (-0,9794; 0,8126)$$

$$\beta_3 = -0,1611 \pm 2,145 \cdot \sqrt{0,0125} = -0,1611 \pm 0,2398 \rightarrow \beta_3 \in (-0,0787; 0,4009)$$

Vậy $\beta_1 \neq 0$, còn $\beta_2 \sim 0$ và $\beta_3 \sim 0$, do đó:

- Hoặc là, mối quan hệ nhân quả $X_2, X_3 \rightarrow Y$ không tồn tại, do đó có thể loại bỏ các yếu tố X_2 và X_3 ra khỏi mô hình nghiên cứu.

- Hoặc là, tồn tại mối quan hệ nhân quả kiểu “đại diện của X_2 ”, “đại diện của X_3 ” $\rightarrow Y$, do đó, cần tiến hành dự đoán và xác định “đại diện của X_2 ” và “đại diện của X_3 ”.

+ Xác định khoảng tin tưởng của μ_0 ứng với $x_1^{(0)} = 26,8$; $x_2^{(0)} = 58$; $x_3^{(0)} = 202$ ở mức xác suất sai lầm $P = 0,05$ như sau:

$$x_0 = (26,8; 58; 202) \rightarrow y_0 = 119,21$$

Vì rằng:

$$\begin{aligned} V[y_0] &= V[\bar{Y}] + (x_1^{(0)} - \bar{x}_1) b_1 + (x_2^{(0)} - \bar{x}_2) b_2 + (x_3^{(0)} - \bar{x}_3) b_3 \\ &= V[\bar{Y}] + (x_1^{(0)} - \bar{x}_1)^2 V[b_1] + (x_2^{(0)} - \bar{x}_2)^2 V[b_2] + \\ &\quad + (x_3^{(0)} - \bar{x}_3)^2 V[b_3] K[b_1 b_3] + 2(x_2^{(0)} - \bar{x}_2)(x_3^{(0)} - \bar{x}_3) K[b_2 b_3] \end{aligned}$$

Do đó:

$$\begin{aligned} S_{y_0}^2 &= \frac{398,83}{18} + (26,8 - 11,94)^2 - 0,2891 + (58 - 42,11)^2 \cdot 0,1745 + \\ &+ (202 - 123)^2 \cdot 0,0125 + 2(26,8 - 11,94)(58 - 42,11)(-0,099) + \\ &+ 2(26,8 - 11,94)(202 - 123)(-0,0004) + 2(58 - 42,11)(202 - 123) \cdot (-0,0132) = 127,24 \end{aligned}$$

$$\mu_0 = 119,21 \pm 2,145 \sqrt{127,24} = 119,21 \pm 24,20 \rightarrow \mu_0 \in (95; 143)$$

Ví dụ 4.3.5(6): Thiết lập mô hình toán học cho mối quan hệ giữa chất lượng nướng của bột (Y) và số giờ nướng (X) ở 170°F như sau:

X_i	0,25	0,50	0,75	1,0	1,5	2,0	3,0	4,0	6,0	8,0
Y_i	93	71	63	54	43	38	29	26	22	20

a) Xử lý số liệu theo mô hình toán học $y = b_0 + bx$ hay $y - \bar{Y} = b(x - \bar{x})$ như sau:

$$n = 10; \sum x_i = 27; \sum x_i^2 = 133,125; \bar{x} = 2,7$$

$$\sum Y_i = 459; \sum Y_i^2 = 26269; \sum x_i Y_i = 783,5; \bar{Y} = 45,9$$

$$Q_x = \sum (x_i - \bar{x})^2 = 133,12 - \frac{27^2}{10} = 60,225$$

$$Q_y = \sum (Y_i - \bar{Y})^2 = 26269 - \frac{459^2}{10} = 5200,9$$

$$Q_{xy} = \sum (x_i - \bar{x})(Y_i - \bar{Y}) = 783,5 - 27 \frac{459^2}{10} = -455,8$$

$$b = \frac{Q_{xy}}{Q_x} = \frac{-455,8}{60,225} = -7,5683$$

$$b_0 = \bar{Y} - b\bar{x} = 45,9 - (-7,5683)(2,7) = 66,3344$$

Vậy: $y = 66,3344 - 7,5683x$

$$Q_R = BG = (-7,5683)(-455,8) = 344,63$$

$$Q_e = Q_y - Q_c = 5200,9 - 3449,63 = 1751,27$$

$$F(1, 8) = \frac{\frac{3449,63}{1751,27}}{\frac{1}{8}} = 15,76; P < 0,005$$

$$r_{xy} = -\sqrt{\frac{3449,63}{5200,9}} = -0,8144$$

b) Xử lý số liệu theo mô hình toán học $\lg y = b_0 + b \lg x$ hay $\lg y - \bar{\lg Y} = b(\lg x - \bar{\lg x})$ như sau:

$\lg x_i$	-0,60206	-0,30103	0,12494	0	0,17609	0,30103	0,47717	0,60206
$\lg Y_i$	1,96848	1,85126	1,79934	1,73239	1,63347	1,57978	1,46240	1,41497

$\lg x_i$	0,77815	0,90309	
$\lg Y_i$	1,34242	1,30103	

$$n = 10; \sum \lg x_i = 2,20951; \sum (\lg x_i)^2 = 2,60154; \bar{\lg x} = 0,22095$$

$$\sum \lg Y_i = 16,08554; \sum (\lg Y_i)^2 = 26,34033;$$

$$\sum (\lg x_i)(\lg Y_i) = 2,56515; \bar{\lg Y_i} = 1,60855$$

$$Q_{\lg x} = 2,60154 - \frac{2,20951^2}{10} = 2,11335$$

$$Q_{\lg y} = 26,34033 - \frac{16,08554^2}{10} = 0,46587$$

$$Q_{\lg x \lg y} = 2,56515 - 2,20951 \cdot \frac{16,08554}{10} = -0,98897$$

$$b = \frac{Q_{lgx \lg y}}{Q_{lgx}} = \frac{-0,98897}{2,11335} = -0,46796$$

$$b_0 = \overline{\lg Y} - b \cdot \overline{\lg x} = 1,60855 - (-0,46796)(0,22095) = 1,71195$$

Vậy: $\lg y = 1,71195 - 0,46796 \lg x$

$$Q_R = BG = 1,71195 - 0,46796 \lg x$$

$$Q_e = Q_{\lg y} - Q_R = 0,46587 - 0,46280 = 0,00307$$

$$F(1, 8) = \frac{\frac{0,46280}{0,00307}}{8} = 1206; P < 0,005$$

$$r_{\lg x \lg y} = -\sqrt{\frac{0,46280}{0,46587}} = -0,9967$$

Bây giờ để biết mô hình toán học nào tốt hơn chúng ta tiến hành so sánh các hệ số tương quan $r_1 = 0,8144$ và $r_2 = 0,9967$ như sau (với giả thuyết $H_0: \rho_1 = \rho_2$):

$$U = \frac{\text{Arth}0,8144 - \text{Arth}0,9967}{\sqrt{\frac{1}{10-3} + \frac{1}{10-3}}} = -3,86; P < 0,001$$

Vậy giả thuyết $H_0: \rho_1 = \rho_2$ bị bác bỏ, nghĩa là $r_1 = 0,8144$ và $r_2 = 0,9967$ là khác nhau (thăng giáng thực sự). Do đó, mô hình toán học dạng $\lg y = 1,71195 - 0,46796 \lg x$ phù hợp với số liệu thực nghiệm hơn.

c) Xử lý số liệu theo mô hình toán học $u = b_0 + bx$ hay $u - \bar{u} = b(x - \bar{x})$ với $u = \frac{1}{y-14}$ như sau:

X_i	0,25	0,50	0,75	1,0	1,5	2,0	3,0	4,0
u_i	0,012658	0,017544	0,020408	0,025	0,034483	0,041667	0,066667	0,083333

X_i	6,0	8,0	
u_i	0,125	0,166667	

$$\sum u_i = 0,593427; \sum u_i^2 = 0,059226; \sum x_i u_i = 2,803970; \bar{u} = 0,059343$$

$$Q_u = 0,059226 - \frac{0,593427^2}{10} = 0,024010$$

$$Q_{xu} = 2,803970 - 27 \cdot \frac{0,593427}{10} = 1,201717$$

$$b = \frac{Q_{xu}}{Q_x} = \frac{1,201717}{60,225} = 0,019954$$

$$b_0 = \bar{u} - b\bar{x} = 0,059343 - (0,019954)(2,7) = 0,005467$$

$$\text{Vậy: } y = \frac{1}{0,005467 + 0,019954x} + 14$$

$$Q_R = BG = (0,019954)(1,201717) = 0,023979$$

$$Q_e = Q_u - Q_R = 0,024010 - 0,023979 = 0,000031$$

$$F(1, 8) = \frac{\frac{0,023979}{0,000031}}{8} = 6188; P < 0,005$$

$$r_{\frac{1}{x} \cdot \frac{1}{y-14}} = \sqrt{\frac{0,023979}{0,024010}} = 0,9987$$

Bây giờ, để biết mô hình toán học nào tốt hơn chúng ta tiến hành so sánh các hệ số tương quan $r_2 = 0,9967$ và $r_3 = 0,9987$ với giả thuyết $H_0: \rho_1 = \rho_2$ như sau:

$$U = \frac{\text{Arth}0,9967 - \text{Arth}0,9987}{\sqrt{\frac{1}{10-3} + \frac{1}{10-3}}} = -0,87; P > 0,3$$

Vậy giả thuyết $H_0: \rho_1 = \rho_2$ được chấp nhận, nghĩa là $r_2 = 0,9967$; và $r_3 = 0,9987$ là như nhau (thăng giáng ngẫu nhiên). Do đó, các mô hình toán học $lgy = 1,71195 - 0,46796x$ và $y = \frac{1}{0,005467 + 0,019954x} + 14$ đều là các mô hình toán học tốt, phù hợp cao đối với các số liệu thực nghiệm thu được.

PHỤ LỤC

Bảng 1. Các giá trị $\chi_p^2(K)$ của phân phối $\chi^2(K)$

K	P												
	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	.00393	.00157	.00982	.00393	.0158	.102	.455	1.32	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.103	.211	.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	.0717	.115	.216	.352	.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	.207	.297	.484	.711	1.06	1.92	6.36	5.39	7.78	9.49	11.1	13.3	14.9
5	.412	.554	.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	.676	.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.78	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.6	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0

Bảng 1 (tiếp theo)

K	P												
	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	20.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7
40	20.7	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	37.7	42.9	49.3	56.3	63.2	67.5	71.4	76.2	79.5
60	35.5	37.5	40.5	43.2	46.5	52.3	59.3	67.0	74.4	79.1	83.3	88.4	92.0

Bảng 2. Các giá trị $F_p(K_1, K_2)$ của phân phối Fisher $F(K_1, K_2)$

K_2	P	K_1								
		1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.4	199.5	215.7	224.6	2302	234.0	236.8	236.9	240.5
	.025	647.8	799.5	864.2	899.6	921.8	937.1	848.2	956.7	963.3
	.010	4052	4999.5	5403	5625	5764	5859	5928	5982	6022
	.005	16211	20000	21615	22500	23056	23437	23715	23925	24091
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	1900	1916	19.25	19.30	19.33	19.35	1937	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.005	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88
4	.100	4.54	4.32	4.19	4.14	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77

Bảng 2 (tiếp theo)

K ₂	P	K ₁								
		1	2	3	4	5	6	7	8	9
6	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77
	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39
	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51
	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34
	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54
	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78

Bảng 2 (tiếp theo)

K.	P	K ₁								
		1	2	3	4	5	6	7	8	9
	010	10.04	7.56	6.55	5.99	5.84	5.39	5.20	5.06	4.94
	005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97
11	100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	005	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54
12	100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20
13	100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
	010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	005	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94
14	100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	050	4.60	3.74	3.34	3.1	2.96	2.85	2.76	2.70	2.65
	025	6.30	4.86	4.24	3.89	3.66	3.50	3.36	3.29	3.21
	010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	005	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72

Bảng 2 (tiếp theo)

K_1											
10	12	15	20	24	30	40	60	120	∞	P	K_2
60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33	.100	1
241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3	.050	
968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018	.025	
6056	6106	6157	6209	6235	6261	6287	6313	6339	6366	.010	
24224	24426	24630	24836	24940	25044	25148	25253	25359	25465	.005	
9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49	.100	2
16.40	19.41	19.43	19.46	19.45	19.46	10.47	19.48	19.49	19.50	.050	
39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50	.025	
99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50	.010	
199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5	.005	
5.23	6.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13	.100	3
8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53	.050	
14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90	.025	
27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13	.010	
43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83	.005	
3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76	.100	4
5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63	.050	
8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26	.025	
14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46	.010	
20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32	.005	
3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10	.100	5
4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36	.050	
6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02	.025	

Bảng 2 (tiếp theo)

K ₁											
10	12	15	20	24	30	40	60	120	σ	P	K ₂
10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02	.010	6
13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14	.005	
2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72	.100	
4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67	.050	
5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85	.025	
7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88	.010	
10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88	.005	7
2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47	.100	
3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23	.050	
4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14	.025	
6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65	.010	
8.38	8.18	7.97	7.75	7.65	7.53	7.42	7.31	7.19	7.08	.005	
2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29	.100	8
3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93	.050	
4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67	.025	
5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86	.010	
7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95	.005	
242	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16	.100	9
3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71	.050	
3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33	.025	
526	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31	.010	
6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19	.005	
2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06	.100	10
2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54	.050	
3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	2.08	.025	
4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91	.010	

Bảng 2 (tiếp theo)

K_1											
10	12	15	20	24	30	40	60	120	00	P	K_2
5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64	.005	11
2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97	.100	
2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40	.050	
3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88	.025	
4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60	.010	
5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.44	4.34	4.23	.005	12
2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90	.100	
2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30	.050	
3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72	.025	
4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36	.010	
5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90	.005	13
2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85	.100	
2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21	.050	
3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60	.025	
4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17	.010	
4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65	.005	14
2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80	.100	
2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13	.050	
3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49	.025	
3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00	.010	
4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.44	.005	

K ₁	P				
		1	2	3	4
15	100	3.07	2.70	2.49	2.36
	.050	4.54	3.68	3.29	3.06
	.025	6.20	4.77	4.15	3.80
	.010	8.68	6.36	5.42	4.89
	.005	10.80	7.70	6.48	5.80
16	100	3.05	2.67	2.46	2.33
	.050	4.49	3.63	3.24	3.01
	.025	6.12	4.89	4.08	3.73
	.010	8.53	6.23	5.29	4.77
	.005	10.58	7.51	6.30	5.64
17	100	3.03	2.64	2.44	2.31
	.050	4.45	3.59	3.20	2.96
	.025	6.04	4.62	4.01	3.66
	.010	8.40	6.11	5.18	4.67
	.005	10.38	7.35	6.16	5.50
18	100	3.01	2.62	2.42	2.29
	.050	4.41	3.55	3.16	2.93
	.025	5.98	4.56	3.95	3.61
	.010	8.29	6.01	5.09	4.58
	.005	10.22	7.21	6.03	5.37
19	100	2.99	2.61	2.40	2.27
	.050	4.38	3.52	3.13	2.90
	.025	5.92	4.51	3.90	3.56

Bảng 2 (tiếp theo)

K_i				
5	6	7	8	9
2.27	2.21	2.18	2.12	2.09
2.90	2.79	2.71	2.64	2.59
3.58	3.41	3.29	3.20	3.12
4.56	4.32	4.14	4.00	3.89
5.37	5.07	4.85	4.67	4.54
2.24	2.18	2.13	2.09	2.06
2.85	2.74	2.65	2.59	2.54
3.50	3.34	3.22	3.12	3.05
4.44	4.20	4.03	3.89	3.78
5.21	4.91	4.69	4.52	4.38
2.22	2.15	2.10	2.06	2.03
2.81	2.70	2.61	2.55	2.49
3.44	3.28	3.16	3.06	2.98
4.34	4.10	3.93	3.79	3.68
5.07	4.78	4.56	4.39	4.25
2.20	2.03	2.08	2.04	2.00
2.77	2.66	2.58	2.51	2.46
3.38	3.22	3.10	3.01	2.93
4.25	4.01	3.84	3.71	3.60
4.96	4.66	4.44	4.28	4.14
2.18	2.11	2.06	2.02	1.98
2.74	2.63	2.54	2.48	2.42
3.33	3.17	3.05	2.96	2.88

Bảng 2 (tiếp theo)

K ₂	P	K ₁								
		1	2	3	4	5	6	7	8	9
20	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.005	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04
	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
21	.010	8.10	5.85	4.94	5.43	4.10	3.87	3.70	3.55	3.46
	.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96
	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	.050	4.32	3.47	3.07	2.34	2.68	2.57	2.49	2.42	2.37
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
22	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	.005	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88
	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
23	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	.005	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81
	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
24	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.005	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75
	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26

Bảng 2 (tiếp theo)

K ₂	P	K ₁								
		1	2	3	4	5	6	7	8	9
25	.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69
	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	.005	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64
	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	.005	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60
	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	.056	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	.005	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56
	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
	.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	.005	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52

Bảng 2. Các giá trị $F_p(K_1, K_2)$ của phân phối Fisher $F(K_1, K_2)$ (tiếp theo)

K_1											
10	12	15	20	24	30	40	60	120	∞	P	K_2
2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76	.100	15
2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07	.050	
3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40	.025	
3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87	.010	
4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26	.005	
2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72	.100	16
2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01	.050	
2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32	.025	
3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75	.010	
4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11	.005	
2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69	.100	17
2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.05	2.01	1.96	.050	
2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25	.025	
3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65	.010	
4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98	.005	
1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66	.100	18
2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92	.050	
2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19	.025	
3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57	.010	
4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.87	.005	
1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63	.100	19
2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88	.050	
2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13	.025	
3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49	.010	
3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78	.005	

Bảng 2 (tiếp theo)

K ₁											
10	12	15	20	24	30	40	60	120	00	P	K ₂
1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61	.100	20
2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84	.050	
2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09	.025	
3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42	.010	
3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69	.005	
1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59	.100	21
2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81	.050	
2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04	.025	
3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36	.010	
3.77	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.61	.005	
1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57	.100	22
2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78	.050	
2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00	.025	
3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31	.010	
3.70	3.54	3.36	3.18	3.08	2.93	2.83	2.77	2.66	2.55	.005	
1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55	.100	23
2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76	.050	
2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97	.025	
3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26	.010	
3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48	.005	
1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53	.100	24
2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73	.050	
2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94	.025	
3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21	.010	
3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43	.005	

Bảng 2 (tiếp theo)

K ₁											P	K ₂
10	12	15	20	24	30	40	60	120	∞			
1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52	.100	.25	
2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71	.050		
2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91	.025		
3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17	.010		
3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38	.005		
1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50	.100	.26	
2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69	.050		
2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.96	.025		
3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13	.010		
3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33	.005		
1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49	.100	.27	
2.20	2.31	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67	.050		
2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85	.025		
3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10	.010		
3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29	.005		
1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48	.100	.28	
2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65	.050		
2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83	.025		
3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06	.010		
3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25	.005		

Bảng 2 (tiếp theo)

K ₂	P	K ₁								
		1	2	3	4	5	6	7	8	9
29	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
	.010	7.60	5.42	4.54	4.04	3.37	3.50	3.33	3.20	3.09
	.005	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48
30	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45
40	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.26	2.53	2.45
	.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	.005	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22
60	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33

Bảng 2 (tiếp theo)

K_2	P	K_1								
		1	2	3	4	5	6	7	8	9
120	.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	.005	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01
	.100	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68
	.050	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
	.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22
∞	.010	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
	.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81
	.100	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63
	.050	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88
	.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11
	.010	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41
	.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62

Bảng 2 (tiếp theo)

K_1											
10	12	15	20	24	30	40	60	120	∞	P	K_2
1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47	.100	29
2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64	.050	
2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81	.025	
3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03	.010	
3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21	.005	
1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46	.100	30
2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62	.050	
2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79	.025	
2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01	.010	
3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18	.005	
1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38	.100	40
2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51	.050	
2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64	.025	
2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80	.010	
3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93	.005	
1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29	.100	60
1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39	.050	
2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48	.025	
2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60	.010	
2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.69	.005	
1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19	.100	120

Bảng 2 (tiếp theo)

K_1											
10	12	15	20	24	30	40		120	∞	P	K_2
1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25	.050	∞
2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31	.025	
2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.31	.010	
2.71	2.54	2.37	2.19	1.99	1.98	1.87	1.75	1.61	1.43	.005	
1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00	.100	
1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00	.050	
2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00	.025	
2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00	.010	
2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00	.005	

Bảng 3. Các giá trị t_p (K) của phân phối student t (K)

K	P								
	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	836.619
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	.718	.906	1.134	1.440	1.934	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.567	2.698	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850

Bảng 3 (tiếp theo)

K	P								
	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
21	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.746
25	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.663	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.422	2.704	3.551
60	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	.674	.842	1.036	1.282	1.645	1.960	2.826	2.576	3.291

Bảng 4. Các giá trị U_p của phân phối bình thường dạng chuẩn $N(0, 1)$

P	0.5	0.4	0.3	0.2	0.1	0.05	0.025	0.020	0.010	0.005	0.001
U_p	.674	.842	1.036	1.282	1.645	1.960	2.241	2.326	2.576	2.807	3.291

Bảng 5 (tiếp theo)

K	P	a													
		2	3	4	5	6	7	8	9	10	12	14	16	18	20
11	.01	4.48	4.73	4.88	4.96	5.06	5.13	5.20	5.24	5.28	5.36	5.42	5.48	5.54	5.55
	.05	3.11	3.27	3.35	3.39	3.43	3.44	3.43	3.46	3.46	3.46	3.46	3.46	3.47	3.48
12	.01	4.39	4.63	4.77	4.86	4.94	5.01	5.06	5.12	5.15	5.24	5.28	5.34	5.38	5.39
	.05	3.08	3.23	3.33	3.36	3.40	3.42	3.44	3.44	3.46	3.46	3.46	3.46	3.47	3.48
13	.01	4.32	4.55	4.68	4.76	4.84	4.92	4.96	5.02	5.07	5.13	5.17	5.22	5.24	5.26
	.05	3.06	3.21	3.30	3.35	3.38	3.41	3.42	3.44	3.45	3.45	3.46	3.46	3.47	3.47
14	.01	4.26	4.48	4.62	4.69	4.74	4.84	4.88	4.94	4.98	5.04	5.08	5.13	5.14	5.15
	.05	3.03	3.18	3.27	3.33	3.37	3.39	3.41	3.42	3.44	3.45	3.46	3.46	3.47	3.47
15	.01	4.21	4.42	4.55	4.63	4.70	4.78	4.83	4.87	4.91	4.96	5.00	5.04	5.06	5.07
	.05	3.01	3.16	3.25	3.31	3.36	3.38	3.40	3.42	3.43	3.44	3.45	3.46	3.47	3.47
16	.01	4.17	4.37	4.50	4.58	4.64	4.72	4.77	4.81	4.84	4.90	4.94	4.97	4.99	5.00
	.05	3.00	3.15	3.23	3.30	3.34	3.37	3.39	3.41	3.43	3.44	3.45	3.46	3.47	3.47
17	.01	4.13	4.34	4.45	4.54	4.60	4.67	4.72	4.76	4.79	4.84	4.88	4.91	4.93	4.94
	.05	2.98	3.13	3.22	3.28	3.33	3.36	3.38	3.40	3.42	3.44	3.45	3.46	3.47	3.47
18	.01	4.10	4.30	4.41	4.50	4.56	4.63	4.68	4.72	4.75	4.80	4.83	4.86	4.88	4.89
	.05	2.87	3.12	3.21	3.27	3.32	3.35	3.37	3.39	3.41	3.43	3.45	3.46	3.47	8.47
19	.01	4.07	4.27	4.38	4.46	4.53	4.59	4.64	4.68	4.71	4.72	4.79	4.82	4.84	4.85
	.05	2.96	3.11	3.19	3.26	3.31	3.35	3.37	3.39	3.41	3.43	3.44	3.46	3.47	3.47
20	.01	4.05	4.24	4.35	4.43	4.50	4.56	4.61	4.64	4.67	4.72	4.76	4.79	4.81	4.82
	.05	2.95	3.10	3.18	3.25	3.30	3.34	3.36	3.38	3.40	3.43	3.44	3.46	3.46	3.47
22	.01	4.02	4.22	4.33	4.40	4.47	4.53	4.58	4.61	4.65	4.69	4.73	4.76	4.78	4.79
	.05	2.93	3.08	3.17	3.24	3.29	3.32	3.35	3.37	3.39	3.42	3.44	3.45	3.46	3.47
	.01	3.99	4.17	4.28	4.36	4.42	4.48	4.53	4.57	4.60	4.65	4.68	4.71	4.74	4.75

K	P							
		2	3	4	5	6	7	8
24	.05	2.92	3.07	3.15	3.22	3.28	3.31	3.34
	.01	3.96	4.14	4.24	4.33	4.39	4.44	4.49
26	.05	2.91	3.06	3.14	3.21	3.27	3.30	3.34
	.01	3.93	4.11	4.21	4.36	4.30	4.41	4.46
28	.03	2.90	3.04	3.13	3.20	3.26	3.30	3.33
	.01	3.91	4.08	4.18	4.28	4.34	4.39	4.43
30	.05	2.89	3.04	3.12	3.20	3.25	3.29	3.32
	.01	3.89	4.06	4.16	4.22	4.32	4.38	4.41
40	.05	2.86	3.01	3.10	3.17	3.22	3.27	3.30
	.01	3.82	3.99	4.10	4.17	4.24	4.30	4.34
60	.05	2.83	2.98	3.08	3.14	3.20	3.24	3.28
	.01	3.76	3.92	4.03	4.12	4.17	4.23	4.27
100	.05	2.80	2.95	3.05	3.12	3.18	3.22	3.26
	.01	3.71	3.86	3.98	4.06	4.11	4.17	4.21
∞	.05	2.77	2.92	3.02	3.09	3.15	3.19	3.23
	.01	3.64	3.80	3.90	3.98	4.04	4.09	4.14

Bảng 5 (tiếp theo)

9	10	12	14	16	18	20
3.37	3.38	3.41	3.44	3.45	3.46	3.47
4.53	4.57	4.62	4.64	4.67	4.70	4.72
3.38	3.38	3.41	3.43	3.45	3.46	3.47
4.50	4.58	4.58	4.62	4.65	4.67	4.69
3.35	3.37	3.40	3.43	3.45	3.46	3.47
4.47	4.51	4.56	4.60	4.62	4.65	4.67
3.35	3.37	3.40	3.43	3.44	3.46	3.47
4.45	4.48	4.54	4.58	4.61	4.63	4.65
3.33	3.35	3.39	3.42	3.44	3.46	3.47
4.37	4.41	4.46	4.51	4.54	4.57	4.59
3.31	3.33	3.37	3.40	3.43	3.45	3.47
4.31	4.34	4.39	4.44	4.47	4.50	4.53
3.29	3.32	3.36	3.40	3.42	3.45	3.47
4.25	4.29	4.35	4.38	4.42	4.45	4.48
3.26	3.29	3.34	3.38	3.41	3.44	3.47
4.17	4.20	4.26	4.31	4.34	4.38	4.41

Bảng 6. Các giá trị $q_p^{(a)}$ (K) dùng trong trắc nghiệm Newman - Keuls và trắc nghiệm Tukey

K	P	a									
		2	3	4	5	6	7	8	9	10	11
5	.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17
	.01	5.70	6.97	7.80	8.42	8.91	9.36	9.67	9.97	10.24	10.48
6	.05	3.46	4.34	4.90	5.31	5.63	5.89	6.12	6.32	6.49	6.65
	.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30
7	.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30
	.01	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55
8	.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05
	.01	4.74	5.63	6.20	6.63	6.96	7.24	7.47	7.68	7.87	8.03
9	.05	3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74	5.87
	.01	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.32	7.49	7.65
10	.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72
	.01	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36
11	.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61
	.01	4.39	5.14	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13
12	.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40	5.51
	.01	4.32	5.04	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94
13	.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43
	.01	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79
14	.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36
	.01	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66
15	.05	3.01	3.67	4.08	4.37	4.60	4.78	4.94	5.08	5.20	5.31
	.01	4.17	4.83	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55

Bảng 6 (tiếp theo)

K	P	a									
		2	3	4	5	6	7	8	9	10	11
16	.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26
	.01	4.13	4.78	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46
17	.05	2.98	3.65	4.02	4.30	4.52	4.71	4.86	4.99	5.11	5.21
	.01	4.10	4.74	5.14	5.34	5.66	5.85	6.01	6.15	6.27	6.38
18	.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17
	.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31
19	.05	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14
	.01	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25
20	.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11
	.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19
24	.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01
	.01	3.96	4.54	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02
30	.05	2.89	3.49	3.84	4.10	4.30	4.46	4.60	4.72	4.83	4.92
	.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85
40	.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74	4.82
	.01	3.82	4.37	4.70	4.93	5.11	5.27	5.39	5.50	5.60	5.69
60	.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73
	.01	3.76	4.28	4.60	4.82	4.99	5.13	5.25	5.36	5.45	5.53
120	.05	2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.48	4.56	4.64
	.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.38
∞	.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55
	.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23

Bảng 7. Các giá trị $q_p^{(a)}$ (K) dùng trong trắc nghiệm Newman-Keuls và trắc nghiệm Tukey

a									P	K
12	13	14	15	16	17	18	19	20		
7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21	.05	5
10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81	11.93	.01	
6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59	.05	6
9.49	9.65	9.81	9.95	10.08	10.21	10.32	10.43	10.54	.01	
6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.09	7.17	.05	7
8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65	.01	
6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87	.05	8
8.18	8.31	8.44	8.65	8.66	8.76	8.85	8.94	9.03	.01	
5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64	.05	9
7.78	7.91	8.03	8.13	8.23	8.32	8.41	8.49	8.57	.01	
5.83	5.93	6.03	6.11	6.20	6.27	6.34	6.40	6.47	.05	10
7.48	7.60	7.71	7.81	7.91	7.99	8.07	8.15	8.22	.01	
8.71	8.81	8.90	8.99	9.06	9.14	9.20	9.26	9.33	.05	11
7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95	.01	
5.62	5.71	5.80	5.88	5.95	6.03	6.09	6.15	6.21	.05	12
7.00	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73	.01	
5.53	5.63	5.71	5.79	5.86	5.93	6.00	6.05	6.11	.05	13
6.90	7.01	7.10	7.19	7.27	7.34	7.42	7.48	7.55	.01	
5.46	5.55	5.64	5.72	5.79	5.85	5.92	5.97	6.03	.05	14
6.77	6.87	6.96	7.05	7.12	7.20	7.27	7.33	7.39	.01	

Bảng 7 (tiếp theo)

u									P	K
12	13	14	15	16	17	18	19	20		
5.40	5.49	5.58	5.65	5.72	5.79	5.85	5.90	5.96	.05	15
6.66	6.76	6.84	6.93	7.00	7.07	7.14	7.20	7.26	.01	
5.35	5.44	5.52	5.59	5.66	5.72	5.79	5.84	5.90	.05	16
6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15	.01	
5.31	5.39	5.47	5.55	5.61	5.68	5.74	5.79	5.84	.05	17
6.48	6.57	6.66	6.73	6.80	6.87	6.94	7.00	7.05	.01	
5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79	.05	18
6.41	6.50	6.58	6.65	6.72	6.79	6.85	6.91	6.96	.01	
5.23	5.32	5.39	5.46	5.53	5.59	5.65	5.70	5.75	.05	19
6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89	.01	
5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71	.05	20
6.29	6.37	6.45	6.52	6.59	6.65	6.71	6.76	6.82	.01	
5.10	5.18	5.25	5.32	5.38	5.44	5.50	5.54	5.59	.05	24
6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61	.01	
5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.48	.05	30
5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41	.01	
4.91	4.98	5.05	5.11	5.16	5.22	5.27	5.31	5.36	.05	40
5.77	5.84	5.90	5.96	6.02	6.07	6.12	6.17	6.21	.01	
4.81	4.88	4.94	5.00	5.06	5.11	5.16	5.20	5.24	.05	60
5.60	5.67	5.73	5.79	5.84	5.89	5.93	5.98	6.02	.01	
4.72	4.78	4.84	4.90	4.95	5.00	5.05	5.09	5.13	.05	120
5.44	5.51	5.56	5.61	5.66	5.71	5.75	5.79	5.83	.01	
4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01	.05	∞
5.29	5.35	5.40	5.48	5.49	5.54	5.57	5.61	5.65	.01	

Bảng 8. Các giá trị $T_{q_p}^{(a)}$ (K) dùng trong trắc nghiệm Dunnett

K	P	a								
		1	2	3	4	5	6	7	8	9
5	.05	2.57	3.03	3.39	3.66	3.88	4.06	4.22	4.36	4.99
	.01	4.03	4.36	5.09	5.44	5.73	5.97	6.18	6.36	6.53
6	.05	2.45	2.86	3.18	3.41	3.60	3.75	3.88	4.00	4.11
	.01	3.71	4.22	4.60	4.88	5.11	5.30	5.47	5.61	5.74
7	.05	2.36	2.75	3.04	3.24	3.41	3.54	3.66	3.76	3.86
	.01	3.50	3.95	4.28	4.52	4.71	4.87	5.01	5.13	5.24
8	.05	2.31	2.67	2.94	3.13	3.28	3.40	3.51	3.60	3.68
	.01	3.36	3.77	4.06	4.27	4.44	4.58	4.70	4.81	4.90
9	.05	2.26	2.61	2.86	3.04	3.18	3.29	3.39	3.48	3.55
	.01	3.25	3.36	3.90	4.09	4.24	4.37	4.48	4.57	4.65
10	.05	2.23	2.57	2.81	2.97	3.11	3.21	3.31	3.39	3.46
	.01	3.17	3.53	3.78	3.95	4.10	4.21	4.31	4.40	4.47
11	.05	2.20	2.53	2.76	2.92	3.05	3.15	3.24	3.31	3.38
	.01	3.11	3.45	3.68	3.85	3.98	4.09	4.18	4.26	4.33
12	.05	2.18	2.50	2.72	2.88	3.00	3.10	3.18	3.25	3.32
	.01	3.05	3.39	3.61	3.76	3.89	3.99	4.08	4.15	4.22
13	.05	2.16	2.48	2.69	2.84	2.96	3.06	3.14	3.21	3.27
	.01	3.01	3.33	3.54	3.69	3.81	3.91	3.99	4.06	4.13
14	.05	2.14	2.46	2.67	2.81	2.93	3.02	3.10	3.17	3.23
	.01	2.98	3.29	3.49	3.64	3.75	3.84	3.92	3.99	4.05

Bảng 8 (tiếp theo)

K	P	a								
		1	2	3	4	5	6	7	8	9
15	.05	2.13	2.44	2.64	2.79	2.90	2.99	3.07	3.13	3.19
	.01	2.95	3.25	3.45	3.59	3.70	3.79	3.86	3.93	3.99
16	.05	2.12	2.24	2.63	2.77	2.88	2.96	3.04	3.10	3.16
	.01	2.92	3.22	3.41	3.55	3.66	3.74	3.82	3.88	3.93
17	.05	2.11	2.41	2.61	2.75	2.85	2.94	3.01	3.08	3.13
	.01	2.90	3.19	3.38	3.51	3.62	3.70	3.77	3.83	3.89
18	.05	2.10	2.40	2.59	2.73	2.84	2.92	2.99	3.05	3.11
	.01	2.88	3.17	3.35	3.48	3.58	3.67	3.74	3.80	3.85
19	.05	2.09	2.39	2.58	2.72	2.82	2.90	2.97	3.04	3.09
	.01	2.86	3.15	3.33	3.46	3.55	3.64	3.70	3.76	3.81
20	.05	2.09	2.38	2.57	2.70	2.81	2.89	2.96	3.02	3.07
	.01	2.85	3.13	3.31	3.344	3.53	3.61	3.67	3.73	3.78
24	.05	2.06	2.35	2.58	2.66	2.76	2.84	2.91	2.96	3.01
	.01	2.80	3.07	3.24	3.36	3.45	3.52	3.58	3.64	3.69
30	.05	2.04	2.32	2.50	2.62	2.72	2.79	2.86	2.91	2.96
	.01	2.75	3.01	3.17	3.28	3.37	3.44	3.50	3.55	3.59
40	.05	2.02	2.29	2.47	2.58	2.67	2.75	2.81	2.86	2.90
	.01	2.70	2.95	3.10	3.21	3.29	3.36	3.41	3.46	3.50
60	.05	2.00	2.27	2.43	2.55	2.63	2.70	2.76	2.81	2.85
	.01	2.66	2.90	3.04	3.14	3.22	3.28	3.33	3.33	3.42
120	.05	1.98	2.24	2.40	2.51	2.59	2.66	2.71	2.76	2.80
	.01	2.62	2.84	2.98	3.08	3.15	3.21	3.25	3.30	3.33
∞	.05	1.96	2.21	2.37	2.47	2.55	2.62	2.67	2.71	2.75
	.01	2.58	2.79	2.92	3.01	3.08	3.14	3.18	3.22	3.25

TÀI LIỆU THAM KHẢO

1. *Bliss, C.I. (1967). Statistics in Biology. McGraw-Hill Book Co, NewYork.*
2. *Cochran, W.G, Cox, G.M. (1964). Experimental designs. John Wiley & Sons, NewYork.*
3. *Fisher, R. A. (1954). Statistical methods for research workers. Oliver & Boyd, Edinburgh, London.*
4. *Nguyễn Ngọc Kiểng (1993). Một số phương pháp cần thiết trong nghiên cứu khoa học. Nhà xuất bản TP. Hồ Chí Minh, TP. Hồ Chí Minh.*
5. *Snedecor, G.W, Cochran, W.G (1971). Statistical methods. The IOWA STATE university press, AMES, IOWA, USA.*
6. *Steel, R. G. D, Torrie, J, H (1960). Principles and procedures of statistics. McGraw-Hill Book Co. NewYork.*
7. *Vukolov, E.A, et al. (1990). Lý thuyết xác suất và thống kê toán học. Nhà xuất bản “khoa học”, Mạc Tư Khoa (Nga văn).*
8. *Williams, E.J (1959). Regression Analysis. John Wiley & Sons, NewYork.*

MỤC LỤC

<i>Lời nói đầu</i>	3
Chương 1: CÁC ĐẠI LƯỢNG THỐNG KÊ	5
1.1. Trung bình	5
1.2. Biến lượng và độ lệch chuẩn	12
1.3. Giao biến lượng (Covariance)	14
Chương 2: CÁC PHÂN PHỐI TRONG THỐNG KÊ HỌC	17
2.1. Phân phối nhị thức, phân phối Poisson và phân phối bình thường	17
2.1.1. Đặc điểm	17
2.1.2. Ứng dụng	25
2.2. Phân phối $\chi^2(K)$	35
2.2.1. Đặc điểm	35
2.2.2. Ứng dụng	41
2.3. Phân phối Student $t(K)$	54
2.3.1. Đặc điểm	54
2.3.2. Ứng dụng	58
2.4. Phân phối Fisher $F(K_1, K_2)$	65
2.4.1. Đặc điểm	65
2.4.2. Ứng dụng	70
Chương 3: SO SÁNH CÁC NGHIỆM THỨC	83
3.1. Khái niệm và các vấn đề liên quan	83
3.1.1. Thuật so sánh trong thống kê học	83
3.1.2. Các phương pháp thu thập số liệu	85
3.1.3. Thuật phân tích biến lượng	87
3.2. So sánh các nghiệm thức dạng tỷ lệ	109
3.2.1. So sánh một tỷ lệ với một tỷ lệ có sẵn	109
3.2.2. So sánh hai tỷ lệ với nhau	111
3.2.3. So sánh nhiều tỷ lệ với nhau	113

3.3. So sánh hai nghiệm thức với nhau	115
3.3.1. Với số liệu không bắt “cặp”	115
3.3.2. Với số liệu bắt “cặp” nhau	121
3.4. So sánh nhiều nghiệm thức với nhau	124
3.4.1. Thuật so sánh đối chiếu	124
3.4.2. Thuật so sánh “phân định”	138
Chương 4: TƯƠNG QUAN, HỒI QUY VÀ PHƯƠNG PHÁP THIẾT LẬP CÁC MÔ HÌNH TOÁN HỌC	155
4.1. Sự tương quan và hồi quy của hai đại lượng ngẫu nhiên	155
4.1.1. Sự tương quan và hồi quy tuyến tính	155
4.1.2. Sự tương quan và hồi quy phi tuyến tính	159
4.1.3. Trường hợp số liệu phân nhóm và mã hóa	161
4.1.4. Thẩm định giá trị của hệ số tương quan	163
4.1.5. Xác lập mô hình hồi quy	168
4.2. Sự tương quan và hồi quy của nhiều đại lượng ngẫu nhiên	172
4.2.1. Sự tương quan riêng phần và đa phần	172
4.2.3. Xác lập mô hình hồi quy	178
4.3. Phương pháp thiết lập các mô hình toán học	178
4.3.1. Mô hình hóa đối tượng nghiên cứu	179
4.3.2. Dự đoán các mô hình toán học	179
4.3.3. Xác định các hệ số của mô hình toán học	182
4.3.4. Thẩm định và đánh giá mức độ của mô hình toán học	187
4.3.5. Một số ví dụ minh họa	197
Phụ lục	213
Tài liệu tham khảo	242

THỐNG KÊ HỌC TRONG NGHIÊN CỨU KHOA HỌC

(Tái bản)

Chịu trách nhiệm xuất bản :

TRINH XUÂN SƠN

Biên tập : HOÀNG NGỌC MINH

Chế bản điện tử: TRẦN THU HOÀI

Sửa bản in : HOÀNG NGỌC MINH

Trình bày bìa : VŨ BÌNH MINH